A Related Works

In the following, we relate our work to recent lines of RLHF research on both theory and practice sides. We also review related works on reward hacking and overoptimization in RLHF.

RLHF: algorithm design. The technique of RLHF [12, 87, 42, 6, 18, 55] has recently demonstrated its great importance in building the state-of-the-art LLMs, including ChatGPT [1], Gemini [61], Claude [2]. In the RLHF pipeline, the LLM is fine-tuned towards maximizing a learned reward model for better alignment [52, 57] with human preference using RL algorithms such as Proximal Policy Optimization (PPO; [51]). Meanwhile, PPO-style algorithm is also known for its instability, sample-inefficiency, and especially, a high demand for proper hyperparameter tuning [22]. This thus casts prohibitive computational cost to make the most effectiveness of PPO-based RLHF methods to align LLMs, especially for the open-source community.

Given that, further research on RLHF has explored various alternatives to PPO-based methods, with the most popular approach being the direct preference optimization method [82, 46], which skips the reward learning phase and directly optimizes the LLM to align it with the human preference. Our practical implementation (RPO) also harnesses the wisdom of reward-LLM equivalence to avoid explicit reward learning followed by PPO training.

Besides the original DPO algorithm [46], ever since it popularizing the direct preference learning style method, variants of the direct preference learning approach are proposed, including but not limited to [34, 5, 71, 59, 28, 74, 44, 26, 50, 32, 80, 58, 68, 30]. Each of them aims to address further challenges of direct preference learning from varying perspectives. Specifically, the algorithm proposed by [44, 26] share similar algorithmic components as RPO proposed in this work. Both work consider SFT style regularization during preference optimization. However, theoretical understanding of how SFT loss can help alignment remains unknown. In contrast, we provide theoretical justifications to the SFT loss as an implicit adversarial regularizer that provably mitigates overoptimization in preference learning.

RLHF: theoretical investigation. Initiated from the literature of dueling bandits and dueling RL [76, 7, 43], recent success of RLHF in fine-tuning LLMs also motivates a long line of research to investigate the theoretical foundations of RLHF under different settings [11, 85, 78, 79, 67, 31, 71, 74, 19, 84], aiming to propose provably sample-efficient algorithms to learn a human-reward-maximizing policy from human preference signals. Our theoretical study of RLHF falls into the paradigm of offline learning from a pre-collected preference dataset, and is mostly related to the work of [85, 78, 31, 71, 74]. In this setup, the main challenge is to address the overoptimization issues due to human reward uncertainty and distributional shifts when only a fixed dataset is available. In the sequel, we compare our work with them in more detail.

Existing theoretical work on provably sample-efficient offline RLHF typically suffers from two drawbacks: they are either restricted to the linear function approximations setting [85, 71] which is far from the practical situations, or are generally unable to be implemented in the LLM experiments. Typically, to encompass the pessimistic principle in the face of uncertainty, the existing literature proposes to return the optimal policy against either an estimated reward model plus a structure-aware reward uncertainty penalty [71] or the most pessimistic reward model inside a confidence region [85, 78]. Both of these two types of method involve intractable components for implementation and needs for additional algorithmic design to approximate the theoretical algorithm in practice. In contrast, our theory works in the context of general function approximations while being friendly to be implemented. Finally, we remark that, while our study focuses on the standard Bradley-Terry model of human preference with general reward function approximations, the work of [74] further considers a general human preference model. But it remains unknown how their algorithms can be efficiently implemented in practice. It serves as an interesting direction to extend our technique to RLHF with general reward model and device new practical algorithms.

Finally, we mention that the algorithm design of RPO is also related to the "pessimism" principle in the standard offline RL literature. It proposes to maintain a pessimistic estimate of the policy values or constrain the policy not to take unseen actions in the data to handle the challenge of the insufficient coverage of the dataset, e.g., [29, 65, 69, 70, 47, 75, 72, 36, 54, 77, 39, 48, 53, 8, 38, 33, 24]. In contrast, we consider the offline RLHF problem and the techniques to obtain the objective of the

RPO algorithm (see Section 4) along with its sample complexity analysis are new and different from these works.

Reward hacking and overoptimization in RLHF for LLM. As is discussed in the introduction, the challenge of reward hacking or overoptimization may prevent the successful alignment of LLMs, degenerating the performance of an LLM because of maximizing an imperfect, overfitted, and misgeneralized proxy reward learned from the finite data [40, 62, 25, 10]. Efforts have been made to mitigate this fundamental issue through the perspective of theory, e.g., [85, 71, 86], and practice, e.g., [15, 21, 41, 81, 49, 56]. Our approach starts from the theoretical insights of handling inherent uncertainty in learning human preference from finite data, while being surprisingly easy to implement.

B Limitations and Future Works

One limitation of the current work is that we focus on the setting of offline RLHF where only a fixed preference dataset is available. Recent RLHF research has shown great potential of using iterative methods for LLM alignment with multiple rounds of preference data collection and tuning [71, 58].

Future works include extending our idea of theoretical algorithm design and analysis to the iterative RLHF setup where further preference data can be collected. Also, since our practical algorithm RPO is a plug-in module that effectively mitigates overoptimization and improves alignment performance, it serves as an exciting direction to combine it with explorative preference data collecting mechanism in iterative RLHF to further boost the performance of LLM alignment.

C Further Discussions

Discussions on Algorithm 1 and Theorem 5.3. We compare our theory with [71] and [78].

Remark C.1 (Comparison with [71]). *Another theoretical work on RLHF [71] explicitly models the KL-regularization between the target policy and the reference policy in the learning objective, referred to as the KL-regularized contextual bandit. This means that their metric becomes the KL-regularized expected reward. In contrast, here we put the KL-regularization as a component of our algorithm design, but we still keep the metric as the expected reward* (2.2)*. Therefore our theory in Section 5.1 directly reveals how the learned policy performs in terms of the expected reward compared to any given target policy (which can be a stochastic policy).*

Remark C.2 (Comparison with [78]). *We remark that in the work of [78], they also mentioned a maximin object similar to* (3.2) *for offline preference-based RL as a complementary to their theoretical algorithm. However, the sample complexity of the maximin-style algorithm they presented is unknown, while we provide finite sample convergence result for Algorithm 1 in Section 5. Furthermore, our objective* (3.2) *features another KL-regularization term, which is essential for the proposal of our new practical algorithm design for aligning LLM in Section 4.*

Discussions on the partial coverage assumption (Assumption 5.2). A sufficient condition to make this partial coverage condition (Assumption 5.2) hold is that the distribution of the offline dataset, which is $\mu_{\mathcal{D}}$, can well cover the joint distribution of $(a^1, a^0) \sim (\pi, \pi^{\mathrm{base}})$. Here to discuss focus on $\pi^{\mathrm{base}} = \pi^{\mathrm{chosen}}$ as we adopted in the experiment part.

First, we clarify that the offline dataset distribution $\mu_{\mathcal{D}}$ is not simply $(a^1, a^0) \sim (\pi^{\mathrm{unchosen}}, \pi^{\mathrm{chosen}})$, since according to our definition (see Section 2) whether a^1 or a^0 is chosen is random and is determined by $y \in 0, 1$ obeying the BT model. Thus, $(a^1, a^0) \sim \mu_{\mathcal{D}}$ can be interpreted as a mixture of $(\pi^{\mathrm{unchosen}}, \pi^{\mathrm{chosen}})$ and $(\pi^{\mathrm{chosen}}, \pi^{\mathrm{unchosen}})$. This mixture probability would not be too small as long as the quality of (a^1, a^0) does not vary too much, i.e., both of them are possible to be chosen, which is the case in practice. As a result, in the offline data distribution $(a^1, a^0) \sim \mu_{\mathcal{D}}$, both a^1 and a^0 partly comes from the chosen distribution π^{chosen}.

Then in order for $\mu_{\mathcal{D}}$ to cover the joint distribution of $(a^1, a^0) \sim (\pi, \pi^{\mathrm{base}})$, it suffices to argue that π^{chosen} can cover the target policy π, which is then reduced back to the traditional coverage condition. Thus our assumption essentially requires that π^{chosen} well covers and only needs to cover the target policy π. This coincides with the spirit of the minimal data assumption in offline RL theory, i.e., the so-called partial coverage condition.

On the relationship between observed chosen probability and reward overoptimization. First, we note that the actions and their chosen probabilities can be interpreted as a proxy of analyzing the underlying (estimated) reward model $\widehat{r}$ due to the representation $\pi_{\widehat{r}}(a|x) \propto \pi^{\mathrm{ref}}(a|x)\exp(\beta^{-1}\widehat{r}(x,a))$. Analyzing the (log) probabilities of the actions can be utilized to detect the mitigation of overoptimization, because according to the representation, an overestimated reward of a poor action would result in a higher probability of choosing this action, and would also cause a decay in the probability of choosing other better actions (since the probabilities are normalized to 1).

To further showcase the ability of RPO to address overoptimization (through the lense of probability), consider the following theoretical example with only one state and three actions [73] where we can track everything clearly. It has three actions a, b, c with $R^\star(a) = 1, R^\star(b) = 0.5, R^\star(c) = 0$. The reference policy $\pi^{\mathrm{ref}}(a) = \pi^{\mathrm{ref}}(b) = 0.4, \pi^{\mathrm{ref}}(c) = 0.1$, and the dataset consists of one data point $\mathcal{D} = (a, b, 1)$ (meaning action a is preferred in the data). Then an ideally solved DPO objective would be π_{DPO} as long as $\pi^{\mathrm{DPO}}(b) = 0$, and the value of $\pi^{\mathrm{DPO}}(a)$ can be arbitrarily chosen in $[0, 1]$. Thus a possible solution to DPO would be $\pi^{\mathrm{DPO}}(a) = 0.5, \pi^{\mathrm{DPO}}(b) = 0$, and by the normalizing condition $\pi^{\mathrm{DPO}}(c) = 0.5$, which is undesirable since the action c has reward $R^\star(c) = 0$. In contrast, solving the RPO objective would additionally require the maximization of $\pi_{\mathrm{RPO}}(a)$ due to the SFT regularization term, and thus the solution is shifted towards $\pi_{\mathrm{RPO}}(a) = 1, \pi_{\mathrm{RPO}}(b) = \pi_{\mathrm{RPO}}(c) = 0$, which is better than the DPO policy. Thus, RPO is able to prevent overoptimization towards poor actions that are less covered by the dataset (action c here), therefore resulting in a better policy.

About the relationships and distinctions between PTX loss in [60] and the SFT loss of RPO. The original PTX loss is an imitation loss calculated on the pretraining data. In contrast, the SFT loss in the RPO objective is an imitation loss calculated on the RLHF dataset. In more specific, our experiments use this SFT loss to imitate the chosen responses in the RLHF dataset. Thus the relationship is that they are both imitation loss which aims to mimic certain data distribution. The distinction is that they are calculated on different data sources. The SFT loss in the RPO objective naturally comes from our theoretical algorithm and provably serves as an important regularization term to mitigate overoptimization in offline RLHF.

About the computational complexity of the SFT loss gradient. According to the paragraph **Practical implementation** in Section 6, RPO adds an additional SFT loss (the log probability of the chosen labels in the preference dataset) on the original DPO loss, where we highlight that the SFT loss is actually an intermediate quantity in the calculation of the DPO loss. Hence, our proposed method does not incur any additional computation overhead compared with the vanilla DPO.

D Proofs for Sample Complexity Analysis

D.1 Proof of Theorem 5.3

Proof of Theorem 5.3. By definition, the suboptimality gap of $\widehat{\pi}$ w.r.t. π is decomposed as following,

$$
\begin{aligned}
&\mathrm{Gap}^{\pi}(\widehat{\pi}) \\
&\quad = \mathbb{E}_{x\sim d_0, a\sim\pi(\cdot|x)}\Big[r^\star(x,a)\Big] - \mathbb{E}_{x\sim d_0, a\sim\widehat{\pi}(\cdot|x)}\Big[r^\star(x,a)\Big] \\
&\quad = \mathbb{E}_{x\sim d_0, a^1\sim\pi(\cdot|x), a^0\sim\pi^{\mathrm{ref}}(\cdot|x)}\Big[r^\star(x,a^1) - r^\star(x,a^0) - \beta\cdot\mathrm{KL}\big(\pi(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] \\
&\qquad\quad - \eta^{-1}\cdot\min_{r\in\mathcal{R}}\left\{\eta\cdot\mathbb{E}_{\substack{x\sim d_0, a^1\sim\widehat{\pi}(\cdot|x),\\ a^0\sim\pi^{\mathrm{base}}(\cdot|x)}}\Big[r(x,a^1) - r(x,a^0) - \beta\cdot\mathrm{KL}\big(\widehat{\pi}(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] + \mathcal{L}_{\mathcal{D}}(r)\right\} \\
&\qquad\quad + \eta^{-1}\cdot\min_{r\in\mathcal{R}}\left\{\eta\cdot\mathbb{E}_{\substack{x\sim d_0, a^1\sim\widehat{\pi}(\cdot|x),\\ a^0\sim\pi^{\mathrm{base}}(\cdot|x)}}\Big[r(x,a^1) - r(x,a^0) - \beta\cdot\mathrm{KL}\big(\widehat{\pi}(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] + \mathcal{L}_{\mathcal{D}}(r)\right\} \\
&\qquad\quad - \mathbb{E}_{x\sim d_0, a^1\sim\widehat{\pi}(\cdot|x), a^0\sim\pi^{\mathrm{base}}(\cdot|x)}\Big[r^\star(x,a^1) - r^\star(x,a^0) - \beta\cdot\mathrm{KL}\big(\widehat{\pi}(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] \\
&\qquad\quad + \beta\cdot\mathbb{E}_{x\sim d_0}\Big[\mathrm{KL}\big(\pi(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big) - \mathrm{KL}\big(\widehat{\pi}(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] \\
&\quad := \text{Term (A)} + \text{Term (B)} + \text{Term (C)},
\end{aligned}
\tag{D.1}
$$

where in the above Term (A), Term (B), and Term (C) are abbreviations for

$$
\begin{aligned}
&\text{Term (A)} \\
&\quad = \mathbb{E}_{x\sim d_0, a^1\sim\pi(\cdot|x), a^0\sim\pi^{\mathrm{base}}(\cdot|x)}\Big[r^\star(x,a^1) - r^\star(x,a^0) - \beta\cdot\mathrm{KL}\big(\pi(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] \\
&\qquad\quad - \eta^{-1}\cdot\min_{r\in\mathcal{R}}\left\{\eta\cdot\mathbb{E}_{\substack{x\sim d_0, a^1\sim\widehat{\pi}(\cdot|x),\\ a^0\sim\pi^{\mathrm{base}}(\cdot|x)}}\Big[r(x,a^1) - r(x,a^0) - \beta\cdot\mathrm{KL}\big(\widehat{\pi}(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] + \mathcal{L}_{\mathcal{D}}(r)\right\},
\end{aligned}
$$

and

$$
\begin{aligned}
&\text{Term (B)} \\
&\quad = \eta^{-1}\cdot\min_{r\in\mathcal{R}}\left\{\eta\cdot\mathbb{E}_{\substack{x\sim d_0, a^1\sim\widehat{\pi}(\cdot|x),\\ a^0\sim\pi^{\mathrm{base}}(\cdot|x)}}\Big[r(x,a^1) - r(x,a^0) - \beta\cdot\mathrm{KL}\big(\widehat{\pi}(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] + \mathcal{L}_{\mathcal{D}}(r)\right\} \\
&\qquad\quad - \mathbb{E}_{x\sim d_0, a^1\sim\widehat{\pi}(\cdot|x), a^0\sim\pi^{\mathrm{base}(\cdot|x)}}\Big[r^\star(x,a^1) - r^\star(x,a^0) - \beta\cdot\mathrm{KL}\big(\widehat{\pi}(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big],
\end{aligned}
$$

and

$$
\text{Term (C)} = \beta\cdot\mathbb{E}_{x\sim d_0}\Big[\mathrm{KL}\big(\pi(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big) - \mathrm{KL}\big(\widehat{\pi}(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big].
$$

In the following, we analyze Term (A) and Term (B) respectively.

Upper bound Term (A). Notice that by the optimality of our choice of policy $\widehat{\pi}$ in (3.2), we have

$$
\begin{aligned}
&\text{Term (A)} \\
&= \mathbb{E}_{x\sim d_0, a^1\sim\pi(\cdot|x), a^0\sim\pi^{\mathrm{base}}(\cdot|x)}\Big[r^\star(x,a^1) - r^\star(x,a^0) - \beta\cdot\mathrm{KL}\big(\pi(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] \qquad \text{(D.2)} \\
&\qquad - \eta^{-1}\cdot\min_{r\in\mathcal{R}}\bigg\{\eta\cdot\mathbb{E}_{\substack{x\sim d_0, a^1\sim\widehat{\pi}(\cdot|x),\\ a^0\sim\pi^{\mathrm{base}}(\cdot|x)}}\Big[r(x,a^1) - r(x,a^0) - \beta\cdot\mathrm{KL}\big(\widehat{\pi}(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] + \mathcal{L}_{\mathcal{D}}(r)\bigg\} \\
&\le \mathbb{E}_{x\sim d_0, a^1\sim\pi(\cdot|x), a^0\sim\pi^{\mathrm{ref}}(\cdot|x)}\Big[r^\star(x,a^1) - r^\star(x,a^0) - \beta\cdot\mathrm{KL}\big(\pi(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] \\
&\qquad - \eta^{-1}\cdot\min_{r\in\mathcal{R}}\bigg\{\eta\cdot\mathbb{E}_{\substack{x\sim d_0, a^1\sim\pi(\cdot|x),\\ a^0\sim\pi^{\mathrm{base}}(\cdot|x)}}\Big[r(x,a^1) - r(x,a^0) - \beta\cdot\mathrm{KL}\big(\pi(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] + \mathcal{L}_{\mathcal{D}}(r)\bigg\} \\
&= \max_{r\in\mathcal{R}}\bigg\{\mathbb{E}_{x\sim d_0, a^1\sim\pi(\cdot|x), a^0\sim\pi^{\mathrm{base}}(\cdot|x)}\Big[\big(r^\star(x,a^1) - r^\star(x,a^0)\big) - \big(r(x,a^1) - r(x,a^0)\big)\Big] - \eta^{-1}\cdot\mathcal{L}_{\mathcal{D}}(r)\bigg\},
\end{aligned}
$$

where in the inequality we apply the optimality of the choice of policy $\widehat{\pi}$ in (3.2).

Upper bound Term (B). For this term, we directly consider the following bound,

$$
\begin{aligned}
&\text{Term (B)} \\
&= \eta^{-1}\cdot\min_{r\in\mathcal{R}}\bigg\{\eta\cdot\mathbb{E}_{\substack{x\sim d_0, a^1\sim\widehat{\pi}(\cdot|x),\\ a^0\sim\pi^{\mathrm{ref}}(\cdot|x)}}\Big[r(x,a^1) - r(x,a^0) - \beta\cdot\mathrm{KL}\big(\widehat{\pi}(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] + \mathcal{L}_{\mathcal{D}}(r)\bigg\} \\
&\qquad - \mathbb{E}_{x\sim d_0, a^1\sim\widehat{\pi}(\cdot|x), a^0\sim\pi^{\mathrm{base}}(\cdot|x)}\Big[r^\star(x,a^1) - r^\star(x,a^0) - \beta\cdot\mathrm{KL}\big(\widehat{\pi}(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] \\
&\le \mathbb{E}_{x\sim d_0, a^1\sim\widehat{\pi}(\cdot|x), a^0\sim\pi^{\mathrm{base}}(\cdot|x)}\Big[r^\star(x,a^1) - r^\star(x,a^0) - \beta\cdot\mathrm{KL}\big(\widehat{\pi}(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] + \eta^{-1}\cdot\mathcal{L}_{\mathcal{D}}(r^\star) \\
&\qquad - \mathbb{E}_{x\sim d_0, a^1\sim\widehat{\pi}(\cdot|x), a^0\sim\pi^{\mathrm{base}}(\cdot|x)}\Big[r^\star(x,a^1) - r^\star(x,a^0) - \beta\cdot\mathrm{KL}\big(\widehat{\pi}(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big] \\
&= \eta^{-1}\cdot\mathcal{L}_{\mathcal{D}}(r^\star), \qquad \text{(D.3)}
\end{aligned}
$$

where in the inequality we apply the fact that $r^\star\in\mathcal{R}$ by Assumption 5.1.

Combining Term (A), Term (B), and Term (C). Now by (D.1), (D.2), and (D.3), we have that

$$
\begin{aligned}
&\mathrm{Gap}_\beta^\pi(\widehat{\pi}) = \text{Term (A)} + \text{Term (B)} + \text{Term (C)} \qquad \text{(D.4)} \\
&\le \max_{r\in\mathcal{R}}\bigg\{\mathbb{E}_{\substack{x\sim d_0, a^1\sim\pi(\cdot|x),\\ a^0\sim\pi^{\mathrm{base}}(\cdot|x)}}\Big[\big(r^\star(x,a^1) - r^\star(x,a^0)\big) - \big(r(x,a^1) - r(x,a^0)\big)\Big] + \eta^{-1}\cdot\Big(\mathcal{L}_{\mathcal{D}}(r^\star) - \mathcal{L}_{\mathcal{D}}(r)\Big)\bigg\} \\
&\qquad + \beta\cdot\mathbb{E}_{x\sim d_0}\Big[\mathrm{KL}\big(\pi(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big) - \mathrm{KL}\big(\widehat{\pi}(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big].
\end{aligned}
$$

In the following, we upper bound the right hand side of (D.4) via relating the MLE loss difference term to the reward difference term through a careful analysis of the preference model. On the one hand, we invoke Lemma D.1 to give an upper bound of the difference of the MLE loss as following, with probability at least $1-\delta$ over random samples and $\varepsilon = (6\cdot(1+e^R)\cdot N)^{-1}$, for any reward model $r\in\mathcal{R}$, it holds that

$$
\begin{aligned}
&\mathcal{L}_{\mathcal{D}}(r^\star) - \mathcal{L}_{\mathcal{D}}(r) \\
&\qquad \le -2\cdot\mathbb{E}_{(x,a^1,a^0)\sim\mu_{\mathcal{D}}(\cdot,\cdot,\cdot)}\Big[D^2_{\mathrm{Hellinger}}\big(\mathbb{P}_{r^\star}(\cdot|x,a^1,a^0)\|\mathbb{P}_r(\cdot|x,a^1,a^0)\big)\Big]
\end{aligned}
$$

$$+ \frac{3}{N} \cdot \log\left(\frac{\mathcal{N}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)}{\delta}\right),$$

where we recall that we use the subscript r in $\mathbb{P}_r$ to emphasize the dependence of the probabilistic model on the reward model. Here $\mathcal{N}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)$ denotes the ε-covering number of the reward model class and R is the upper bound on the reward functionss (Assumption 5.1). Now to facilitate the calculation, we lower bound the Hellinger distance by total variation (TV) distance as

$$D^2_{\text{Hellinger}}\big(\mathbb{P}_{r^\star}(\cdot|x,a^1,a^0)\|\mathbb{P}_r(\cdot|x,a^1,a^0)\big) \geq D^2_{\text{TV}}\big(\mathbb{P}_{r^\star}(\cdot|x,a^1,a^0)\|\mathbb{P}_r(\cdot|x,a^1,a^0)\big),$$

By the expression of the probability model $\mathbb{P}_r$, we can further write the TV distance above as

$$\begin{aligned}
&D_{\text{TV}}\big(\mathbb{P}_{r^\star}(\cdot|x,a^1,a^0)\|\mathbb{P}_r(\cdot|x,a^1,a^0)\big) \\
&\quad= \frac{1}{2} \cdot \Big|\sigma\big(r^\star(x,a^1) - r^\star(x,a^0)\big) - \sigma\big(r(x,a^1) - r(x,a^0)\big)\Big| \\
&\qquad + \frac{1}{2} \cdot \Big|\sigma\big(r^\star(x,a^0) - r^\star(x,a^1)\big) - \sigma\big(r(x,a^0) - r(x,a^1)\big)\Big| \\
&\quad= \Big|\sigma\big(r^\star(x,a^1) - r^\star(x,a^0)\big) - \sigma\big(r(x,a^1) - r(x,a^0)\big)\Big|,
\end{aligned} \tag{D.5}$$

where in the second equality we use the fact that $\sigma(-z) = 1 - \sigma(z)$. Now by Lemma D.2 and the condition that $r(x,a) \in [0,R]$ for any $(x,a,r) \in \mathcal{X} \times \mathcal{A} \times \mathcal{R}$ (Assumption 5.1), we know that

$$\begin{aligned}
&\Big|\sigma\big(r^\star(x,a^1) - r^\star(x,a^0)\big) - \sigma\big(r(x,a^1) - r(x,a^0)\big)\Big| \\
&\qquad \geq \kappa \cdot \Big|\big(r^\star(x,a^1) - r^\star(x,a^0)\big) - \big(r(x,a^1) - r(x,a^0)\big)\Big|,
\end{aligned}$$

where $\kappa = 1/(1+\exp(R))^2$. As a result, the difference of the MLE loss is upper bounded by

$$\begin{aligned}
&\mathcal{L}_{\mathcal{D}}(r^\star) - \mathcal{L}_{\mathcal{D}}(r) \\
&\quad \leq -2\kappa^2 \cdot \mathbb{E}_{(x,a^1,a^0)\sim\mu_{\mathcal{D}}(\cdot,\cdot,\cdot)}\bigg[\Big|\big(r^\star(x,a^1) - r^\star(x,a^0)\big) - \big(r(x,a^1) - r(x,a^0)\big)\Big|^2\bigg] \\
&\qquad + \frac{3}{N} \cdot \log\left(\frac{\mathcal{N}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)}{\delta}\right).
\end{aligned} \tag{D.6}$$

On the other hand, the reward difference term in (D.4), which is evaluated on actions from π and π^{base}, can be related to the reward difference evaluated on the data distribution $\mu_{\mathcal{D}}$ via Assumption 5.2, i.e.,

$$\begin{aligned}
&\mathbb{E}_{x\sim d_0, a^1\sim\pi(\cdot|x), a^0\sim\pi^{\text{base}}(\cdot|x)}\Big[\big(r^\star(x,a^1) - r^\star(x,a^0)\big) - \big(r(x,a^1) - r(x,a^0)\big)\Big] \qquad \text{(D.7)} \\
&\quad \leq C_{\mu_{\mathcal{D}}}(\mathcal{R};\pi,\pi^{\text{base}})\sqrt{\mathbb{E}_{(x,a^1,a^0)\sim\mu_{\mathcal{D}}}\bigg[\Big|\big(r^\star(x,a^1) - r^\star(x,a^0)\big) - \big(r(x,a^1) - r(x,a^0)\big)\Big|^2\bigg]}.
\end{aligned}$$

Finally, combining (D.6), (D.7), and (D.4), denoting

$$\Delta_r := \sqrt{\mathbb{E}_{(x,a^1,a^0)\sim\mu_{\mathcal{D}}}\bigg[\Big|\big(r^\star(x,a^1) - r^\star(x,a^0)\big) - \big(r(x,a^1) - r(x,a^0)\big)\Big|^2\bigg]},$$

we have that

$$\begin{aligned}
\text{Gap}^\pi(\widehat{\pi}) &\leq \max_{r\in\mathcal{R}}\Big\{C_{\mu_{\mathcal{D}}}(\mathcal{R};\pi,\pi^{\text{base}}) \cdot \Delta_r - 2\eta^{-1}\kappa^2 \cdot \Delta_r^2\Big\} + \frac{3}{\eta N} \cdot \log\left(\frac{\mathcal{N}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)}{\delta}\right) \\
&\qquad + \beta \cdot \mathbb{E}_{x\sim d_0}\Big[\text{KL}\big(\pi(\cdot|x)\|\pi^{\text{ref}}(\cdot|x)\big) - \text{KL}\big(\widehat{\pi}(\cdot|x)\|\pi^{\text{ref}}(\cdot|x)\big)\Big] \\
&\leq \frac{\big(C_{\mu_{\mathcal{D}}}(\mathcal{R};\pi,\pi^{\text{base}})\big)^2\eta}{8\kappa^2} + \frac{3}{\eta N} \cdot \log\left(\frac{\mathcal{N}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)}{\delta}\right) \\
&\qquad + \beta \cdot \mathbb{E}_{x\sim d_0}\Big[\text{KL}\big(\pi(\cdot|x)\|\pi^{\text{ref}}(\cdot|x)\big)\Big],
\end{aligned}$$

where in the second inequality we use that fact that $az - bz^2 \le a^2/(4b)$ for any $z \in \mathbb{R}$ and that KL-divergence is non-negative. Consequently, with the choice of

$$\eta = 2\sqrt{6} \cdot \sqrt{\frac{\log\left(\mathcal{N}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)/\delta\right)}{N}}, \quad \beta = \frac{1}{\sqrt{N}}, \quad \kappa = \frac{1}{(1+\exp(R))^2},$$

we conclude that with probability at least $1-\delta$ and $\varepsilon = (6 \cdot (1+e^R) \cdot N)^{-1}$,

$$\begin{aligned}
&\mathrm{Gap}^\pi(\widehat{\pi}) \\
&\quad \le \frac{\sqrt{6}\big(1+\exp(R)\big)^2 \left(\left(C_{\mu_\mathcal{D}}(\mathcal{R}; \pi, \pi^{\mathrm{base}})\right)^2 + 1\right)\iota + 4\mathbb{E}_{x\sim d_0}\left[\mathrm{KL}\big(\pi(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\right]}{4\sqrt{N}},
\end{aligned}$$

where we denote $\iota = \sqrt{\log\left(\mathcal{N}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)/\delta\right)}$ with $\varepsilon = (6\cdot(1+e^R)\cdot N)^{-1}$. This finishes the proof of Theorem 5.3. ☐

D.2 Technical Lemmas

Lemma D.1 (Uniform concentration)**.** *Consider the MLE loss (3.1) and define the approximation error as $\varepsilon = (6\cdot(1+e^R)\cdot N)^{-1}$ where R is the upper bound on the reward functions (Assumption 5.2). Suppose that the reward model class $\mathcal{R}$ has a finite ε-covering number $\mathcal{N}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty) < \infty$. Then for any $\delta < 1/e$ it holds with probability at least $1-\delta$ that*

$$\begin{aligned}
&\mathcal{L}_\mathcal{D}(r^\star) - \mathcal{L}_\mathcal{D}(r) \\
&\quad \le -2 \cdot \mathbb{E}_{(x,a^1,a^0)\sim\mu_\mathcal{D}(\cdot,\cdot,\cdot)}\left[D^2_{\mathrm{Hellinger}}\big(\mathbb{P}_{r^\star}(\cdot|x,a^1,a^0)\|\mathbb{P}_r(\cdot|x,a^1,a^0)\big)\right] \\
&\qquad + \frac{3}{N}\cdot\log\left(\frac{\mathcal{N}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)}{\delta}\right).
\end{aligned}$$

Proof of Lemma D.1. For notational simplicity, we use $\mathcal{C}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)$ to denote an ε-cover of the reward model class $\mathcal{R}$ under the $\|\cdot\|_\infty$-norm. It holds that $\mathcal{N}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty) = |\mathcal{C}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)|$. First we invoke Proposition 5.3 of [37] to obtain a uniform concentration over the finite set of ε-cover $\mathcal{C}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)$. Specifically, with probability at least $1-\delta$, for any $r \in \mathcal{C}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)$,

$$\begin{aligned}
&\mathcal{L}_\mathcal{D}(r^\star) - \mathcal{L}_\mathcal{D}(r) \\
&\quad \le -2 \cdot \mathbb{E}_{(x,a^1,a^0)\sim\mu_\mathcal{D}(\cdot,\cdot,\cdot)}\left[D^2_{\mathrm{Hellinger}}\big(\mathbb{P}_{r^\star}(\cdot|x,a^1,a^0)\|\mathbb{P}_r(\cdot|x,a^1,a^0)\big)\right] \\
&\qquad + \frac{2}{N}\cdot\log\left(\frac{\mathcal{N}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)}{\delta}\right).
\end{aligned} \tag{D.8}$$

Now for any reward model $r \in \mathcal{R}$, we take a $r^\dagger \in \mathcal{C}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)$ satisfying $\|r - r^\dagger\|_\infty \le \varepsilon$. We have

$$\begin{aligned}
&\mathcal{L}_\mathcal{D}(r^\star) - \mathcal{L}_\mathcal{D}(r) \\
&\quad = \mathcal{L}_\mathcal{D}(r^\star) - \mathcal{L}_\mathcal{D}(r^\dagger) + \mathcal{L}_\mathcal{D}(r^\dagger) - \mathcal{L}_\mathcal{D}(r) \\
&\quad \le -2 \cdot \mathbb{E}_{(x,a^1,a^0)\sim\mu_\mathcal{D}(\cdot,\cdot,\cdot)}\left[D^2_{\mathrm{Hellinger}}\big(\mathbb{P}_{r^\star}(\cdot|x,a^1,a^0)\|\mathbb{P}_{r^\dagger}(\cdot|x,a^1,a^0)\big)\right] \\
&\qquad + \frac{2}{N}\cdot\log\left(\frac{\mathcal{N}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)}{\delta}\right) + \mathcal{L}_\mathcal{D}(r^\dagger) - \mathcal{L}_\mathcal{D}(r) \\
&\quad \le -2 \cdot \mathbb{E}_{(x,a^1,a^0)\sim\mu_\mathcal{D}(\cdot,\cdot,\cdot)}\left[D^2_{\mathrm{Hellinger}}\big(\mathbb{P}_{r^\star}(\cdot|x,a^1,a^0)\|\mathbb{P}_r(\cdot|x,a^1,a^0)\big)\right] \\
&\qquad + \frac{2}{N}\cdot\log\left(\frac{\mathcal{N}_\varepsilon(\mathcal{R}, \|\cdot\|_\infty)}{\delta}\right) + \mathcal{L}_\mathcal{D}(r^\dagger) - \mathcal{L}_\mathcal{D}(r) \\
&\qquad\quad + 4 \cdot \mathbb{E}_{(x,a^1,a^0)\sim\mu_\mathcal{D}(\cdot,\cdot,\cdot)}\left[D^2_{\mathrm{Hellinger}}\big(\mathbb{P}_{r^\dagger}(\cdot|x,a^1,a^0)\|\mathbb{P}_r(\cdot|x,a^1,a^0)\big)\right],
\end{aligned} \tag{D.9}$$

where in the fir inequality we use (D.8) for $r^\dagger$ and in the second inequality we utilize the triangular inequality for Hellinger distance. Therefore, it remains to upper bound the approximation error induced by $r^\dagger$. On the one hand, by the definition of $\mathcal{L}_\mathcal{D}$ in (3.1), we have that

$$\mathcal{L}_\mathcal{D}(r^\dagger) - \mathcal{L}_\mathcal{D}(r)$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\left(\frac{\sigma\big(r(x_i,a_i^1)-r(x_i,a_i^0)\big)}{\sigma\big(r^\dagger(x_i,a_i^1)-r^\dagger(x_i,a_i^0)\big)}\right) \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}(1-y_i)\cdot \log\left(\frac{\sigma\big(r(x_i,a_i^0)-r(x_i,a_i^1)\big)}{\sigma\big(r^\dagger(x_i,a_i^0)-r^\dagger(x_i,a_i^1)\big)}\right).
\end{aligned}
$$

Use the inequality that $\log(x) \le x-1$, we can further upper bound $\mathcal{L}_\mathcal{D}(r^\dagger)-\mathcal{L}_\mathcal{D}(r)$ by

$$
\begin{aligned}
&\mathcal{L}_\mathcal{D}(r^\dagger)-\mathcal{L}_\mathcal{D}(r) \\
&\quad \le \frac{1}{N}\sum_{i=1}^{N} y_i \cdot \frac{\sigma\big(r(x_i,a_i^1)-r(x_i,a_i^0)\big)-\sigma\big(r^\dagger(x_i,a_i^1)-r^\dagger(x_i,a_i^0)\big)}{\sigma\big(r^\dagger(x_i,a_i^1)-r^\dagger(x_i,a_i^0)\big)} \\
&\qquad + \frac{1}{N}\sum_{i=1}^{N}(1-y_i)\cdot \frac{\sigma\big(r(x_i,a_i^0)-r(x_i,a_i^1)\big)-\sigma\big(r^\dagger(x_i,a_i^0)-r^\dagger(x_i,a_i^1)\big)}{\sigma\big(r^\dagger(x_i,a_i^0)-r^\dagger(x_i,a_i^1)\big)}.
\end{aligned}
$$

Now since $\|r^\dagger - r\|_\infty \le \varepsilon$ and $r^\dagger \in [0,R]$, invoking Lemma D.2, we can derive that

$$
\begin{aligned}
\mathcal{L}_\mathcal{D}(r^\dagger)-\mathcal{L}_\mathcal{D}(r) &\le \frac{1}{N}\sum_{i=1}^{N} \frac{\big|\big(r(x_i,a_i^1)-r(x_i,a_i^0)\big)-\big(r^\dagger(x_i,a_i^1)-r^\dagger(x_i,a_i^0)\big)\big|}{(1+e^R)^{-1}} \\
&\quad + \frac{1}{N}\sum_{i=1}^{N} \frac{\big|\big(r(x_i,a_i^0)-r(x_i,a_i^1)\big)-\big(r^\dagger(x_i,a_i^0)-r^\dagger(x_i,a_i^1)\big)\big|}{(1+e^R)^{-1}} \\
&\le 4\cdot \|r^\dagger - r\|_\infty \cdot (1+e^R) \le 4\varepsilon\cdot(1+e^R). \qquad \text{(D.10)}
\end{aligned}
$$

On the other hand, we upper bound the hellinger distance between $\mathbb{P}_r$ and $\mathbb{P}_{r^\dagger}$, for any $(x,a^1,a^0) \in \mathcal{X}\times\mathcal{A}\times\mathcal{A}$,

$$
\begin{aligned}
&D^2_{\text{Hellinger}}\big(\mathbb{P}_{r^\dagger}(\cdot|x,a^1,a^0)\|\mathbb{P}_r(\cdot|x,a^1,a^0)\big) \\
&\quad \le D_{\text{TV}}\big(\mathbb{P}_{r^\dagger}(\cdot|x,a^1,a^0)\|\mathbb{P}_r(\cdot|x,a^1,a^0)\big) \\
&\quad = \Big|\sigma\big(r^\dagger(x,a^1)-r^\dagger(x,a^0)\big)-\sigma\big(r(x,a^1)-r(x,a^0)\big)\Big| \\
&\quad \le \Big|\big(r^\dagger(x,a^1)-r^\dagger(x,a^0)\big)-\big(r(x,a^1)-r(x,a^0)\big)\Big| \\
&\quad \le 2\cdot\|r^\dagger - r\|_\infty \le 2\varepsilon, \qquad \text{(D.11)}
\end{aligned}
$$

where the first inequality uses the fact that $D^2_{\text{Hellinger}} \le D_{\text{TV}}$, the equality uses the same argument as (D.5), and the second inequality applies Lemma D.2. Finally, combining (D.9), (D.10), and (D.11), we conclude that

$$
\begin{aligned}
\mathcal{L}_\mathcal{D}(r^\star)-\mathcal{L}_\mathcal{D}(r) &\le -2\cdot \mathbb{E}_{(x,a^1,a^0)\sim\mu_\mathcal{D}(\cdot,\cdot,\cdot)}\Big[D^2_{\text{Hellinger}}\big(\mathbb{P}_{r^\star}(\cdot|x,a^1,a^0)\|\mathbb{P}_r(\cdot|x,a^1,a^0)\big)\Big] \\
&\quad + \frac{2}{N}\cdot\log\left(\frac{\mathcal{N}_\varepsilon(\mathcal{R},\|\cdot\|_\infty)}{\delta}\right) + 6\varepsilon\cdot(1+e^R).
\end{aligned}
$$

By taking the approximation error $\varepsilon = (6\cdot(1+e^R)\cdot N)^{-1}$, we conclude that for $\delta < e^{-1}$, with probability at least $1-\delta$, for any $r\in\mathcal{R}$, it holds that

$$
\begin{aligned}
&\mathcal{L}_\mathcal{D}(r^\star)-\mathcal{L}_\mathcal{D}(r) \\
&\quad \le -2\cdot \mathbb{E}_{(x,a^1,a^0)\sim\mu_\mathcal{D}(\cdot,\cdot,\cdot)}\Big[D^2_{\text{Hellinger}}\big(\mathbb{P}_{r^\star}(\cdot|x,a^1,a^0)\|\mathbb{P}_r(\cdot|x,a^1,a^0)\big)\Big] \\
&\qquad + \frac{3}{N}\cdot\log\left(\frac{\mathcal{N}_\varepsilon(\mathcal{R},\|\cdot\|_\infty)}{\delta}\right).
\end{aligned}
$$

This completes the proof of Lemma D.1. $\square$

Lemma D.2 (Sigmoid function). *For any real numbers $z_1, z_2 \in [-R,R]$, it holds that*

$$
\kappa\cdot|z_1 - z_2| \le |\sigma(z_1)-\sigma(z_2)| \le |z_1 - z_2|,
$$

where the constant $\kappa = 1/(1+\exp(R))^2$.

Proof of Lemma D.2. Since the sigmoid function $\sigma(\cdot)$ is differentiable, we know that for any $z_1, z_2 \in [-R, R]$, there exists some $\xi(z_1, z_2) \in [-R, R]$ such that

$$\sigma(z_1) - \sigma(z_2) = \sigma'\big(\xi(z_1, z_2)\big) \cdot (z_1 - z_2).$$

Notice that $\sigma'(z) = \sigma(z) \cdot (1 - \sigma(z))$, we can obtain that

$$\begin{aligned}
1 \geq \sigma'\big(\xi(z_1, z_2)\big) &= \sigma\big(\xi(z_1, z_2)\big) \cdot \Big(1 - \sigma\big(\xi(z_1, z_2)\big)\Big) \\
&= \frac{1}{1 + \exp(\xi(z_1, z_2))} \cdot \left(1 - \frac{1}{1 + \exp(\xi(z_1, z_2))}\right) \\
&\geq \frac{1}{1 + \exp(R)} \cdot \left(1 - \frac{1}{1 + \exp(-R)}\right) \\
&= \frac{1}{(1 + \exp(R))^2}.
\end{aligned}$$

This completes the proof of Lemma D.2. □

E Proofs for Equivalence between Maximin and Minimax Objectives

E.1 Proof of Theorem 5.6

Proof of Theorem 5.6. Consider denoting an auxiliary policy $\widehat{\pi}$ as

$$\widehat{\pi} \in \operatorname*{argmax}_{\pi \in \Pi} \min_{r \in \mathcal{R}} \phi(\pi, r). \tag{E.1}$$

By the definition of $\widehat{r}$ and $\widehat{\pi}$, the duality gap of $(\widehat{r}, \widehat{\pi})$, defined as

$$\mathrm{Dual}(\widehat{r}, \widehat{\pi}) := \max_{\pi \in \Pi} \phi(\pi, \widehat{r}) - \min_{r \in \mathcal{R}} \phi(\widehat{\pi}, r)$$

is zero. This is because the following deduction,

$$\begin{aligned}
\mathrm{Dual}(\widehat{r}, \widehat{\pi}) &= \left(\max_{\pi \in \Pi} \phi(\pi, \widehat{r}) - \min_{r \in \mathcal{R}} \max_{\pi \in \Pi} \phi(\pi, r)\right) \\
&\qquad + \left(\max_{\pi \in \Pi} \min_{r \in \mathcal{R}} \phi(\pi, r) - \min_{r \in \mathcal{R}} \phi(\widehat{\pi}, r)\right) \\
&= 0,
\end{aligned} \tag{E.2}$$

where in the first equality we apply Lemma E.1 that the minimax objective and the maximin objective are equivalent, and the last equality applies the definition of $\widehat{r}$ and $\widehat{\pi}$ respectively. Note that we can rewrite the duality gap as following

$$\mathrm{Dual}(\widehat{r}, \widehat{\pi}) = \left(\max_{\pi \in \Pi} \phi(\pi, \widehat{r}) + \phi(\widehat{\pi}, \widehat{r})\right) - \left(\phi(\widehat{\pi}, \widehat{r}) - \min_{r \in \mathcal{R}} \phi(\widehat{\pi}, r)\right). \tag{E.3}$$

Combining (E.2) and (E.3), we can conclude that

$$\max_{\pi \in \Pi} \phi(\pi, \widehat{r}) = \phi(\widehat{\pi}, \widehat{r}) \quad \Rightarrow \quad \widehat{\pi} \in \operatorname*{argmax}_{\pi \in \Pi} \phi(\widehat{r}, \pi). \tag{E.4}$$

Now comparing what $\pi_{\widehat{r}}$ and $\widehat{\pi}$ satisfy in (5.4) and (E.4) respectively, invoking Lemma E.3 that the maximizer of $\phi(\cdot, r)$ given any $r \in \mathcal{R}$ is unique on the support of d_0, we can conclude that

$$\pi_{\widehat{r}}(\cdot|x) = \widehat{\pi}(\cdot|x), \quad \forall x \in \mathrm{Supp}(d_0). \tag{E.5}$$

Therefore, by (E.1) and (E.5), and the fact that $\phi(\pi, r)$ depends on π only through its value on the support of d_0, we can conclude that

$$\pi_{\widehat{r}} \in \operatorname*{argmax}_{\pi \in \Pi} \min_{r \in \mathcal{R}} \phi(\pi, r).$$

This finishes the proof of Theorem 5.6. □

E.2 Auxiliary Lemmas

Lemma E.1 (Equivalence of maximin and minimax objectives)**.** *For the policy class Π defined in* (2.3) *and the reward model class $\mathcal{R}$ satisfying Assumption 5.5, it holds that the maximin objective is equivalent to the minimax objective, i.e.,*

$$\max_{\pi\in\Pi}\min_{r\in\mathcal{R}}\phi(\pi,r)=\min_{r\in\mathcal{R}}\max_{\pi\in\Pi}\phi(\pi,r).$$

Proof of Lemma E.1. The foundation of this result is a minimax theorem given by [23] (Lemma E.2). In our setting, the policy class Π is a nonempty set, and the reward model class $\mathcal{R}$ is a nonempty compact Hausdorff space. Furthermore, by our choice of the policy class Π in (2.3), Π is a convex set. Meanwhile, the function ϕ is a concave function of $\pi\in\Pi$ since the dependence on π is linear terms plus a negative KL term (concave). Finally, by our assumption, the function ϕ is convex-like on the reward model class $\mathcal{R}$ and is also continuous on $\mathcal{R}$. As a result, all the conditions of Lemma E.2 are satisfied and the minimax theorem holds in our problem setup, finishing the proof of Lemma E.1. ☐

Lemma E.2 (Minimax theorem [23])**.** *Let $\mathcal{X}$ be a nonempty set (not necessarily topologized) and $\mathcal{Y}$ be a nonempty compact topological space. Let $f:\mathcal{X}\times\mathcal{Y}\mapsto\mathbb{R}$ be lower semicontinuous on $\mathcal{Y}$. Suppose that f is concave-like on $\mathcal{X}$ and convex-like on $\mathcal{Y}$, i.e., for any $x_1,x_2\in\mathcal{X}$, $\alpha\in[0,1]$, there exists $x_3\in\mathcal{X}$ such that*

$$f(x_3,\cdot)\geq\alpha\cdot f(x_1,\cdot)+(1-\alpha)\cdot f(x_2,\cdot)\ \textit{on}\ \mathcal{Y},$$

and for any $y_1,y_2\in\mathcal{Y}$, $\beta\in[0,1]$, there exists $y_3\in\mathcal{Y}$ such that

$$f(\cdot,y_3)\leq\beta\cdot f(\cdot,y_1)+(1-\beta)\cdot f(\cdot,y_2)\ \textit{on}\ \mathcal{Y}.$$

Then the following equation holds,

$$\max_{x\in\mathcal{X}}\min_{y\in\mathcal{Y}}f(x,y)=\min_{y\in\mathcal{Y}}\max_{x\in\mathcal{X}}f(x,y).$$

Lemma E.3 (Unique maximizer of ϕ)**.** *Consider the function ϕ defined as*

$$\begin{aligned}\phi(\pi,r):=\eta\cdot\mathbb{E}_{x\sim d_0,a^1\sim\pi(\cdot|x),a^0\sim\pi^{\mathrm{base}}(\cdot|x)}\Big[r(x,a^1)-r(x,a^0)-\beta\cdot D_{\mathrm{KL}}\big(\pi(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big]\\+\mathcal{L}_{\mathcal{D}}(r).\end{aligned}$$

Then given any $r\in\mathcal{R}$, the maximimzer of $\phi(\cdot,r)$ is unique on the support of d_0.

Proof of Lemma E.3. Given any $r\in\mathcal{R}$, consider that

$$\begin{aligned}&\max_{\pi\in\Pi}\phi(\pi,r)\\&\quad=\eta\cdot\max_{\pi\in\Pi}\Big\{\mathbb{E}_{x\sim d_0,a^1\sim\pi(\cdot|x)}\Big[r(x,a^1)-\beta\cdot D_{\mathrm{KL}}\big(\pi(\cdot|x)\|\pi^{\mathrm{ref}}(\cdot|x)\big)\Big]\Big\}\\&\quad=\eta\cdot\max_{\pi\in\Pi}\left\{C_r-\beta\cdot\mathbb{E}_{x\sim d_0}\left[D_{\mathrm{KL}}\left(\pi(\cdot|x)\middle\|\frac{\pi^{\mathrm{ref}}(\cdot|x)\cdot\exp(\beta^{-1}\cdot r(x,\cdot))}{\int_{a'\in\mathcal{A}}\mathrm{d}\pi^{\mathrm{ref}}(a'|x)\cdot\exp(\beta^{-1}\cdot r(x,a'))}\right)\right]\right\},\end{aligned}$$

where

$$C_r=\mathbb{E}_{x\sim d_0}\left[\beta\cdot\log\left(\int_{a\in\mathcal{A}}\mathrm{d}\pi^{\mathrm{ref}}(a|x)\cdot\exp\left(\beta^{-1}\cdot r(x,a)\right)\right)\right]$$

is a constant independent of π. Therefore, the maximizer of $\phi(\cdot,r)$ on the support of d_0 must equal to

$$\pi_r(\cdot|x)=\frac{\pi^{\mathrm{ref}}(\cdot|x)\cdot\exp(\beta^{-1}\cdot r(x,\cdot))}{\int_{a'\in\mathcal{A}}\mathrm{d}\pi^{\mathrm{ref}}(a'|x)\cdot\exp(\beta^{-1}\cdot r(x,a'))},$$

which completes the proof of Lemma E.3. ☐

F Additional Details on Experiments

F.1 Training Details

We train the gemma series models with 8 NVIDIA A6000 GPUs and the beta series models with 8 NVIDIA A100 GPUs, where they are all GPT-like models with around 7 billion parameters. It takes around three hours to train a beta series model and five hours to train a gemma one. Our codebase is adapted from the Alignment Handbook [63]. By comparing the validation loss on the test split (not used for later evaluation), we select the hyperparameter η of both RPO (beta) and RPO (gemma) to be 0.005. We list the remaining training configurations in Table 3, which are recommended by the Alignment Handbook.

Configuration	Beta Series	Gemma Series
learning rate	5.0e-7	5.0e-7
learning scheduler type	cosine	cosine
warmup ratio	1.0	1.0
batch size	128	128
gradient accumulation	2	16
batch size per device	8	1
training epoch	1	2
β	0.01	0.05
optimizer	adamw torch	adamw torch
seed	42	42
precision	bfloat16	bfloat16

Table 3: Training configurations for beta series and gemma series models in this paper.

F.2 Evaluation Details

GPT-4 evaluation on the test split. We use the following prompts to guide GPT-4 to annotate the preferences among win, lose, and tie (we denote them by A, B, and C, respectively).

> **Prompts:** Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: [[A]] if assistant A is better, [[B]] if assistant B is better, and [[C]] for a tie. [Instruction] instruction [The Start of Assistant A's Answer] {*answer A*} [The End of Assistant A's Answer] [The Start of Assistant B's Answer] {*answer B*} [The End of Assistant B's Answer]

Here, we replace {*answer A*} and {*answer B*} with the answers of two models. Since GPT annotation has shown to prefer the answer in the first position [66], we randomly exchange the positions between two answers during the evaluation to ensure a fair comparison.

Benchmark evaluation. We use the default configuration for the evaluations on MT-Bench[2] and AlpacaEval 2.0[3]. By default, the annotator of MT-Bench is the *latest version* of GPT-4. The default annotator and the competitor model are both GPT-4 (Preview 11/06). We only need to manually import the proper chat template that formats the training dataset, which are shown as follows.

[2] `https://github.com/lm-sys/FastChat/tree/main/fastchat/llm_judge`
[3] `https://github.com/tatsu-lab/alpaca_eval/tree/main`

Chat Template for Beta Series: <|system|></s><|user|>
{*instruction*}</s>
<|assistant|>

Chat Template for Gemma Series: <bos> <|im_start|>user
{*instruction*}<|im_end|>
<|im_start|>assistant

F.3 Additional Results on Experiments

In this section, we provide the additional results to show the performance gain for RPO (beta) in MT-Bench and RPO (gemma) in AlpacaEval 2.0. We report the pairwise win rates in Tables 4, 5, and 6 to analyze their performance gaps, where all the annotation configurations are the same in Table 2. Results show that RPO still exceeds DPO in the metric of the pairwise win rates on the benchmarks for both beta series and gemma series.

win rate (%)	RPO (beta)	Ref. (beta)	DPO (beta)
RPO (beta)	50.00	**83.75**	**57.81**
Ref. (beta)	16.25	50.00	21.25
DPO (beta)	78.75	42.19	50.00

Table 4: Pairwise win rates (left vs. right) for beta series models on MT-Benchmark.

win rate (%)	RPO (beta)	Ref. (beta)	DPO (beta)
RPO(beta)	50.00	**80.13**	**52.02**
Ref.(beta)	19.87	50.00	20.61
DPO (beta)	47.98	79.39	50.00

Table 5: Pairwise win rates (left vs. right) for gemma series models on AlpacaEval 2.0.

win rate (%)	RPO (beta)	Ref. (beta)	DPO (beta)
RPO (beta)	50.00	**64.93**	**51.33**
Ref. (beta)	35.07	50.00	36.44
DPO (beta)	48.67	64.56	50.00

Table 6: Pairwise Length-Control (LC) win rates (left vs. right) for gemma series models on AlpacaEval 2.0.

G Experiments on Math, Reasoning, and Coding Tasks

G.1 Experimental Details

To provide a more comprehensive analysis of the trained LLM, we introduce more benchmarks on the math, reasoning, and coding tasks for evaluations. Specifically, we choose the Grade School Math 8K (GSM8K), AI2 Reasoning Challenge (ARC), and Mostly Basic Python Programming (MBPP) to measure math, reasoning, and coding abilities, respectively. In this section, we focus on the gemma series for the experiments. We do not use chain-of-thought or few shots in all the benchmarks. We compare the greedy decoding result (pass @1) on the MBPP benchmark.

Model Name	GSM8K (%)	ARC		MBPP (Pass @1)	
		Easy (%)	Challenge (%)	Normal (%)	Plus (%)
RPO	**49.9**	**79.1**	49.8	54.2	**46.3**
DPO	45.3	75.7	**50.0**	54.2	43.9
Ref.	45.4	75.0	45.8	50.3	44.2
`zephyr-gemma-7b`	47.3	77.6	48.6	**54.5**	44.7

Table 8: Results on GSM8K, ARC, and MBPP. Here, `zephyr-gemma-7b` is the officially released models trained by DPO and Ref. denotes the reference model `zephyr-7b-gemma-sft` used for our training. RPO and DPO are trained with the OpenRLHF codebase [27] and we average the SFT loss regularizer in RPO by the number of tokens of the chosen response. We do not use chain-of-thought or few shots in all the benchmarks. We compare the greedy decoding result (pass @1) for MBPP.

Here we use the OpenRLHF codebase [27] to implement a new variant of RPO, where the SFT loss regularizer is averaged by the number of tokens of the chosen labels, that is, $(\log \pi_\theta(a_{\text{cho}}|x))/|a_{\text{cho}}|$. Such a variant balances the weight of the averaged SFT loss regularizer between the shorter chosen response and the longer one. We set the coefficient for the SFT loss regularizer as 0.2. We use 8 NVIDIA A100 GPUs for the training and evaluation. The remaining hyperparameters are in Table 7.

Configuration	Gemma Series
learning rate	5.0e-7
learning scheduler type	cosine with a minimum learning rate
batch size	128
gradient accumulation	8
batch size per device	2
training epoch	2
β	0.5
optimizer	adamw torch
seed	42
precision	bfloat16

Table 7: Training configurations for DPO and RPO for the experiments in Appendix G.

G.2 Experimental Results

Table 8 demonstrates that our proposed method still outperforms or performs equally to the vanilla DPO on these benchmarks of math, reasoning, and coding, which verifies the effectiveness of our proposed method.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We support all the claims made in the abstract and the introduction sections by our theory and experiment sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Please see the discussion of limitation of the work in Appendix B.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide the complete and accurate proof in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the detailed training and evaluation configurations in Appendix F.1 and F.2 for reproducibility. We also submit the codes in the supplementary for the purpose of reproducibility.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We submit the codes in the supplementary. All the datasets, reference models, and benchmarks used in this paper are open accessed.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see the dataset choice and the detailed training configurations in Section 6 and Appendix F.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to the budget of time and expense, we do not report the error bar, which is also common in many RLHF literature [46, 34, 71, 64]. However, we report all the training configurations and the random seed to ensure reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the information in Appendix F.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed and followed the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The scope of our work is not to publish a new released model but to analyze the overoptimization pheromone in RLHF both theoretically and empirically.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We respect all the licenses and terms of codes, models, and datasets used in this paper. We also properly cite their creators in this paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.