Doubly Robust Alignment for Large Language Models

Anonymous Authors[1]

Abstract

This paper studies reinforcement learning from human feedback (RLHF) for aligning large language models with human preferences. While RLHF has demonstrated promising results, many algorithms are highly sensitive to misspecifications in the underlying preference model (e.g., the Bradley-Terry model), the reference policy, or the reward function, resulting in undesirable fine-tuning. To address model misspecification, we propose a doubly robust preference optimization algorithm that remains consistent when either the preference model or the reference policy is correctly specified (without requiring both). Our proposal demonstrates superior and more robust performance than state-of-the-art algorithms, both in theory and in practice.

1. Introduction

Recent advances in large language models (LLMs) have revolutionized various natural language processing tasks, ranging from text generation to human-AI conversation and more complex reasoning tasks (Brown et al., 2020; Touvron et al., 2023; Guo et al., 2025). These models are typically trained in two stages. In the pre-training stage, LLMs learn general linguistic patterns and commonsense knowledge from vast, unlabeled text data through autoregressive next-token prediction. However, pretrained models face a critical objective mismatch: while they are optimized for token prediction, real-world deployment requires alignment with complex human values such as helpfulness, honesty and harmlessness (Askell et al., 2021). This mismatch calls for an additional post-training stage, aiming at better aligning these pre-trained models with human preference.

The paper studies reinforcement learning from human feedback (RLHF), a post-training paradigm that adapts pre-trained models through reinforcement learning (RL, Sutton et al., 2018). The RLHF literature has rapidly expanded in recent years, where existing algorithms can be broadly categorized as reward-based or preference-based (Section 2 for a review). While demonstrating remarkable success in domains including robotics control, video games, and LLMs fine-tuning (see e.g., Christiano et al., 2017; Ziegler et al., 2019; Bai et al., 2022; Bakker et al., 2022; Ouyang et al., 2022), they often suffer from various model misspecifications (see also Table 1 for a summary):

1. **Preference model misspecification**. Most reward-based algorithms rely on the Bradley-Terry (BT, Bradley & Terry, 1952) preference model (see Equation 1). However, this model entails various unrealistic assumptions on human preference, including transitivity, context-independence and perfect relationality, which are likely violated based on empirical evidence (May, 1954; Tversky, 1969; Gardner, 1970; Agranov & Ortoleva, 2015; Michaud et al., 2020; Milano et al., 2021; Lindner & El-Assady, 2022). While some preference-based algorithms impose more general preference model (GPM) assumptions (see e.g., Zhang et al., 2024d), their effectiveness still depends on correct model specification.
2. **Reward model misspecification**. Under the BT model assumption, classical reward-based algorithms first estimate the reward function from human preference data and then apply RL algorithms such as the proximal policy optimization (**?**)PPO,][]schulman2017proximal to derive the optimal policy. However, policy learning through RL is highly sensitive to the estimated reward. Misspecifying the reward can lead to reward hacking (Skalse et al., 2022; Laidlaw et al., 2024) and misguide policy learning (Kaufmann et al., 2023; Zheng et al., 2023; Chen et al., 2024).
3. **Reference policy misspecification**. To alleviate misspecification of the reward, recent algorithms based on direct preference optimization (DPO, Rafailov et al., 2023) propose to express the reward in closed form using the reference policy for policy learning. However, these algorithms are sensitive to the specification of reference policy (Liu et al., 2024b; Gorbatovski et al., 2024; Xu et al., 2024).

Drawing from doubly robust estimation methods in econo-

---

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the ICML 2025 Workshop on Reliable and Responsible Foundation Models. Do not distribute.

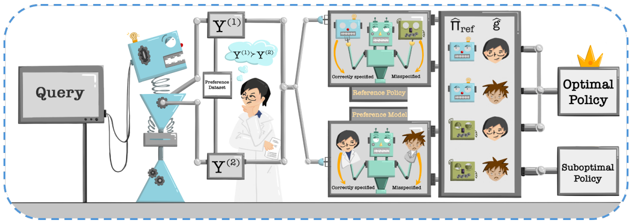


Figure 1. A visualization of our proposed preference optimization algorithm. $\widehat{\pi}_{\text{ref}}$ denotes the specified reference policy whereas $\widehat{g}$ denotes the specified preference model. Our proposal is doubly robust in that it requires correct specification of either the reference policy, or the preference model.

metrics and RL (see Section 2 for a literature review), we introduce a novel RLHF algorithm that is robust to model misspecification and statistically efficient; see Figure 1 for a visualization of our algorithm. Our major contributions are summarized as follows:

- We propose a robust and efficient estimator for preference evaluation, i.e., evaluating the probability of a target policy being preferred over the reference policy. The proposed preference estimator achieves two desirable properties: (i) *double robustness* (Corollary 3) – it converges to the true preference probability when either the preference model or the reference policy is correctly specified, and (ii) *semi-parametric efficiency* (Corollary 4) – it attains the smallest mean squared error (MSE) among all regular and asymptotically linear estimators (Newey, 1990; Tsiatis, 2006).
- Leveraging this preference estimator, we further develop a preference optimization algorithm for LLM fine-tuning. The proposed algorithm maintains *double robustness* (Corollary (6)) and remains consistent even when the BT model assumption is violated (Theorem 5). Meanwhile, when the BT model assumption holds, its suboptimality gap is *less sensitive* to the reward model and reference policy compared to PPO- or DPO-based algorithms, and is likely smaller than that of both types of algorithms (Theorem 7).

2. Related Works

Our work is closely related to reward- and preference-based RLHF algorithms, as well as doubly robust (DR) methods. We discuss these related works below.

Reward-based RLHF. Reward-based algorithms assume the existence of a latent utility or reward function that determines human preferences, estimate the reward function from the data and apply RL for policy learning. Recent research has focused on addressing practical challenges such as reward hacking and model-collapse. These issues arise due to PPO's sensitivity to reward specification, gradient clipping thresholds, and the tuning parameter controlling KullbackLeibler (KL)-divergence regularization (Engstrom et al., 2019; Zheng et al., 2023; Xiao et al., 2024). Existing approaches to these challenges fall into three categories: (i) The first category focuses on improving the reward learning algorithm to obtain more accurate reward functions (Li et al., 2023b; Chan et al., 2024; Gao et al., 2024; Liu et al., 2024a; Fu et al., 2025; Xiao et al., 2025; Ye et al., 2025). (ii) The second category develops better policy learning algorithms using the estimated reward function (Wu et al., 2024a; Zhang et al., 2024a; Shao et al., 2024; Hu, 2025; Liu et al., 2025b; Yu et al., 2025; Yuan et al., 2025). (iii) The third category is DPO-based, which bypasses reward learning entirely and directly optimizes policies under the BT model assumption (Zhao et al., 2022; Wang et al., 2023a; Ethayarajh et al., 2024; Song et al., 2024; Tang et al., 2024).

Unlike many of these algorithms, our proposal does not rely on the BT model assumption, and is more robust to the misspecification of reward or reference policy when the BT model holds.

Preference-based RLHF. Preference-based algorithms do not assume the existence of a latent reward function at all; instead, they search the optimal policy that maximizes the alignment with human preferences (see e.g., Azar et al., 2024). In particular, there is a growing line of research that adopts the Nash learning from human feedback (NLHF, Munos et al., 2023) framework, which formulates the alignment problem as a two-player constant-sum game and solves for policies that achieve the Nash equilibrium (Calandriello et al., 2024; Rosset et al., 2024; Swamy et al., 2024; Wu et al., 2024b; Ye et al., 2024; Zhang et al., 2024c; Liu et al., 2025a). Beyond NLHF, (Wang et al., 2023b) develops a Bayesian approach for alignment, whereas (Hong et al., 2024) and (Zhang et al., 2024d) propose energy-based and general preference models to relax the BT model assumption.

Our proposal belongs to this class of preference-based methods. In particular, the proposed algorithm is most closely related to the identity preference optimization (IPO) algorithm proposed by Azar et al. (2024), as both maximizing the same objective function in the population level (see Section 4 for the objective). However, unlike IPO, our proposed method is robust to misspecifications of the reference policy. Similarly, compared to Zhang et al. (2024d), the proposed algorithm is more robust to the misspecification of the preference model. Finally, our work differs from NLHF in its primary focus: we study robust and *statistically efficient* preference estimation from data, rather than developing *computationally efficient* algorithms to solve the Nash equilibrium.

Doubly robust methods. DR has been extensively studied in statistics, econometrics and machine learning. These methods originate from the missing data and causal inference literature (see e.g., Robins et al., 1994; Scharfstein et al., 1999). To illustrate these methods, consider the fundamental causal inference problem of estimating the average treatment effect (ATE) – the difference in the mean outcome between a newly-developed treatment policy and a baseline policy for a given patient population. DR first estimates two models from the data: (i) a propensity score model (similar to the reference policy in LLMs) that characterizes the treatment assignment mechanism and (ii) an outcome regression model (similar to the reward function) that specifies the conditional mean function of a patient's outcome. It then employs both models to construct the ATE estimator, whose consistency requires only one of the models to be correct. Furthermore, when both models are correct, the resulting estimator is semiparametrically efficient (Bang & Robins, 2005). The favorable statistical properties of these methods have led to extensive follow-up research (see e.g., Tan, 2010; Tsiatis et al., 2011; Imai & Ratkovic, 2014; Vermeulen & Vansteelandt, 2015; Kennedy et al., 2017; Robins et al., 2017; Wager & Athey, 2018; Wang & Tchetgen Tchetgen, 2018; Künzel et al., 2019; Oprescu et al., 2019; Shi et al., 2019; Fulcher et al., 2020; Farrell et al., 2021; Nie & Wager, 2021; Li et al., 2022; Cui et al., 2023; Han et al., 2023; Kennedy, 2023; Li et al., 2023a; Cui et al., 2024; Wang & Shah, 2024; Wang & Han, 2024). A seminal extension appears in Chernozhukov et al. (2018), which proposes to learn both the propensity score and outcome regression models using machine learning methods to deal with complex data structures with high-dimensional covariates, texts or images.

Beyond treatment effect estimation in causal inference, doubly robust methods have been widely applied to a broad range of other problems, including the estimation and evaluation of optimal (dynamic) treatment regimes (Robins, 2004; Zhang et al., 2012; 2013; Schulte et al., 2015; Luedtke & Van Der Laan, 2016; Fan et al., 2017; Jiang et al., 2017; Song et al., 2017; Shi et al., 2018), conditional independence testing (Zhang et al., 2019; Shah & Peters, 2020; Shi et al., 2021; Quinzan et al., 2023; Zhang et al., 2024b), offline policy learning (Dudík et al., 2014; Kallus & Uehara, 2020b; Uehara et al., 2020b; Liao et al., 2022; Shi et al., 2024a) and off-policy evaluation (OPE, Jiang & Li, 2016; Thomas & Brunskill, 2016; Farajtabar et al., 2018; Kallus & Zhou, 2018; Bibaut et al., 2019; Tang et al., 2020; Kallus & Uehara, 2020a; Su et al., 2020; Uehara et al., 2020a; Shi et al., 2022; Xie et al., 2023; Xu et al., 2023; Cao & Zhou, 2024; Shi et al., 2024b).

However, none of the aforementioned works considers the application of fine-tuning LLMs – a gap we aim to bridge by connecting these two vibrant research areas.

3. RLHF Preliminaries: Data, Modeling and Baseline Algorithms

Data generating process. Assume we are given a dataset $\mathcal{D}$, consisting of n i.i.d. tuples of the form $(X, Y^{(1)}, Y^{(2)}, Z)$. Each of these tuples is generated as follows: Given a prompt X, two independent responses $(Y^{(1)}, Y^{(2)})$ are generated under a reference policy π_{ref} such that $Y^{(1)}, Y^{(2)} \sim \pi_{\text{ref}}(\bullet|X)$. These data $(X, Y^{(1)}, Y^{(2)})$ are then shown to a human expert, who provides a binary preference $Z = \mathbb{I}(Y^{(1)} \succ Y^{(2)})$ where $Y^{(1)} \succ Y^{(2)}$ indicates that the first response is preferred, and $\mathbb{I}(\bullet)$ denotes the indicator function. Additionally, let g^* denote the preference function such that $g^*(X, Y^{(1)}, Y^{(2)}) = \mathbb{P}(Y^{(1)} \succ Y^{(2)}|X)$ determines the probability of $Y^{(1)}$ being favored over $Y^{(2)}$ conditional on X.

We remark that the reference policy π_{ref} is not always known. For instance, the responses might be generated by an LLM different from the target model that we wish to fine-tune (Bai et al., 2022). Furthermore, the responses might be produced by a heterogeneous set of models rather than a single model (Stiennon et al., 2020).

BT model. As commented in Section 2, most existing reward-based RLHF algorithms impose the BT model assumption, which requires the preference function g^* to take the following form,

$$g^*(x, y^{(1)}, y^{(2)}) = \sigma(r^*(y^{(1)}, x) - r^*(y^{(2)}, x)), \tag{1}$$

where r^* denotes some underlying reward function that measures how well a response answers a given prompt, and σ denotes the sigmoid function. As commented in the introduction, this assumption is likely violated due to the inherent intransitivity, inconsistency and stochasticity in human preference.

Assuming (1) holds, the goal is to learn an optimal policy π^* that maximizes the expected reward

$$J(\pi) = \mathbb{E}[\mathbb{E}_{y\sim\pi(\bullet|X)} r^*(y, X)], \tag{2}$$

among all policies π. Here, the outer expectation is taken with respect to the prompt distribution, whereas the inner expectation is taken with respect to the response generated by a given policy π.

We next introduce two types of baseline algorithms – PPO-based and DPO-based – for learning π^*. Both approaches operate under Assumption (1).

PPO-based approaches. PPO-based algorithms proceed in two steps. In the first step, they compute an estimated

Table 1. Robustness of different algorithms to model misspecification. Our algorithm is denoted by DRPO, short for doubly robust preference optimization.

		Robust to misspecified:	preference model	reward model	reference policy
RLHF	Reward-based	PPO-type	✗	✗	✓
		DPO-type	✗	✓	✗
	Preference-based	IPO (Azar et al., 2024)	✓	-	✗
		GPM (Zhang et al., 2024d)	✗	-	✓
		DRPO	✓	✓	✓

reward function $\widehat{r}$ using maximum likelihood estimation or empirical risk minimization. In the second step, they learn π^* by maximizing

$$\mathbb{E}_{X\sim\mathcal{D},\, y\sim\pi(\bullet|X)}\left[\widehat{r}(y,X)\right] - \beta\, D_{\mathrm{KL}}\left[\pi(y\mid X)\,\|\,\pi_{\mathrm{ref}}(y\mid X)\right], \tag{3}$$

over $\pi\in\Pi$ (e.g., a transformer-based policy class), where the expectation is taken over prompts X from the empirical data distribution and responses y from a target policy π, D_{KL} denotes the KL divergence measure between the target and reference policies, and the tuning parameter $\beta>0$ controls the degree to which π is allowed to deviate from π_{ref}. The KL regularization term in (3) encourages the learned policy to stay close to π_{ref}, in order to mitigate overfitting and prevent the learned policy from collapsing to a narrow set of high-reward responses (Zheng et al., 2023).

DPO-based approaches. DPO-based algorithms are motivated by the fact that the argmax to (3) (denoted by $\widehat{\pi}$) can be represented in closed-form using the estimated reward $\widehat{r}$. This in turn yields the following closed-form expression for $\widehat{r}$,

$$\widehat{r}(y,x) = \beta\log\left(\frac{\widehat{\pi}(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right) - C(x), \tag{4}$$

for some response-independent function $C(x)$ that will cancel out in pairwise comparisons. As such, instead of solving $\widehat{\pi}$ in two steps, DPO-based approaches directly parameterize the reward via Equation (4) and compute $\widehat{\pi}$ in a single step – for example, by maximizing the likelihood of the human preference data under the BT model.

To conclude this section, we note that, as shown in Equation (3), the optimal policy computed by PPO can be highly sensitive to the estimated reward function $\widehat{r}$. While DPO-based approaches eliminate this dependence, Equation (4) reveals that their optimization relies on the specification of the reference policy π_{ref}. Due to these sensitivities, even under the idealized setting where the BT model holds, both PPO- and DPO-based algorithms can underperform our proposed algorithm, which is inherently more robust to misspecification in both $\widehat{r}$ and π_{ref}. We provide theoretical justification in Section 5 and empirical validation in Section 6.

4. Double Robust Preference Evaluation and Optimization

This section introduces the proposed doubly robust approach; see Figure 1 for a visualization. Different from these reward-based algorithms discussed in Section 3, we adopt a preference-based approach that searches the optimal policy by maximizing its total preference. Specifically, given a target policy π, its *total preference* over the reference policy (Azar et al., 2024) is defined by

$$p^*(\pi) := \mathbb{P}(\pi\succ\pi_{\mathrm{ref}}) = \mathbb{E}[\mathbb{E}_{y\sim\pi(\bullet|X),\, y'\sim\pi_{\mathrm{ref}}(\bullet|X)}\, g^*(X,y,y')],$$

where we recall that g^* denotes the preference function $\mathbb{P}(y>y'|X)$, and the outer expectation is taken with respect to the prompt distribution. As both $Y^{(1)}$ and $Y^{(2)}$ are generated under π_{ref}, we have

$$p^*(\pi) = \frac{1}{2}\sum_{a=1}^{2}\mathbb{E}[\mathbb{E}_{y\sim\pi(\bullet|X)}\, g^*(X,y,Y^{(a)})]. \tag{5}$$

For preference evaluation, our goal is to accurately estimate $p^*(\pi)$ for a given target policy π from the dataset $\mathcal{D}$. In the following, we first introduce two baseline estimators: a direct method (DM) estimator and an importance sampling (IS) estimator, where the names are borrowed from the OPE literature (see e.g., Uehara et al., 2022). We next introduce our proposed DR estimator, which combines both DM and IS for efficient and robust preference evaluation.

DM estimator. The direct method estimator is motivated by (5). It proceeds by first estimating g^* and then plugging the estimated g^* (denoted by $\widehat{g}$) into (5) to construct the estimator,

$$\widehat{p}_{\mathrm{DM}}(\pi) = \frac{1}{2}\mathbb{E}_{X\sim\mathcal{D},y\sim\pi(\bullet|X)}[\widehat{g}(X,y,Y^{(1)})+\widehat{g}(X,y,Y^{(2)})], \tag{6}$$

where X is drawn from the empirical data distribution, y is drawn from π and the expectation can be approximated using Monte Carlo sampling.

When an external preference model is available, it can be used directly as $\widehat{g}$, as in (Munos et al., 2023). Otherwise, g^* can be estimated from the data $\mathcal{D}$. For instance, under the

BT model assumption, one can estimate the reward function r^* and plug the estimator into (1) to derive $\widehat{g}$. Alternatively, one can employ more general preference models that do not rely on the BT model.

IS estimator. The second baseline estimator is the IS estimator, which is motivated by the following lemma that expresses $p^*(\pi)$ using the IS ratio $w(y,x) = \pi(y|x)/\pi_{\text{ref}}(y|x)$.

Lemma 1. *Assume* $w(y,x) < \infty$ *for any* x, y. *Then* $p^*(\pi) = \frac{1}{2}\mathbb{E}[w(Y^{(1)},X)Z + w(Y^{(2)},X)(1-Z)]$.

The proof of Lemma 1 is straightforward. It follows directly from the symmetry of pairwise comparisons where the preference can be equivalently expressed using either $g^*(X,y,y')$ or $1-g^*(X,y',y)$, and an application of the change-of-measure theorem (see Appendix A.1).

Based on this identity, we define the following IS estimator:

$$\begin{aligned}\widehat{p}_{\text{IS}}(\pi) = \quad & \tfrac{1}{2}\mathbb{E}_{(X,Y^{(1)},Y^{(2)},Z)\sim\mathcal{D}}\Big[\tfrac{\pi(Y^{(1)}|X)}{\widehat{\pi}_{\text{ref}}(Y^{(1)}|X)}Z \\ & + \tfrac{\pi(Y^{(2)}|X)}{\widehat{\pi}_{\text{ref}}(Y^{(2)}|X)}(1-Z)\Big], \qquad (7)\end{aligned}$$

where $\widehat{\pi}_{\text{ref}}$ denotes an estimated reference policy. If π_{ref} is known, we can directly use the oracle reference policy. Otherwise, for some external datasets (e.g., Bai et al., 2022), well-trained reference models are available and can be used as $\widehat{\pi}_{\text{ref}}$. Finally, when no such external model is available and π_{ref} is unknown, we estimate it from the observed data tuples $(X,Y^{(1)},Y^{(2)})$ using supervised fine-tuning (SFT).

DR estimator. A closer look at Equations (6) and (7) reveals that the DM and IS estimators' consistencies depend crucially on the correct specification of the preference function and the reference policy. We next introduce our proposed DR estimator, which is more robust to misspecifications in these models. It relies on the following estimating function $\psi(X,Y^{(1)},Y^{(2)},Z;\pi,\widehat{\pi}_{\text{ref}},\widehat{g})$, defined as

$$\begin{aligned}&\frac{1}{2}\sum_{a=1}^{2}\mathbb{E}_{y\sim\pi(\bullet|X)}[\widehat{g}(X,y,Y^{(a)})] \\ &+\frac{1}{2}\sum_{a=1}^{2}(-1)^{a-1}\frac{\pi(Y^{(a)}|X)}{\widehat{\pi}_{\text{ref}}(Y^{(a)}|X)}[Z-\widehat{g}(X,Y^{(1)},Y^{(2)})].\end{aligned} \qquad (8)$$

By definition, this estimating function contains two terms: (i) the first term is essentially the estimating function of the DM estimator in (6), and (ii) the second term is an augmentation term, which is similar to IS in (7), but with the observed preference Z replaced by its residual $Z-\widehat{g}(X,Y^{(1)},Y^{(2)})$. The purpose of introducing the additional augmentation term is to correct for the bias introduced by misspecification of the preference model in the DM estimator. This leads to our DR estimator,

$$\widehat{p}_{\text{DR}}(\pi) = \mathbb{E}_{(X,Y^{(1)},Y^{(2)},Z)\sim\mathcal{D}}\psi(X,Y^{(1)},Y^{(2)},Z;\pi,\widehat{\pi}_{\text{ref}},\widehat{g}). \qquad (9)$$

Similar to the DR estimator in the bandit setting (Dudík et al., 2014), (9) is reduced to the DM estimator when setting $\widehat{g}$ to zero, and the IS estimator when setting the IS ratio $\pi/\widehat{\pi}_{\text{ref}}$ to zero. However, as shown in (8), a key different from those bandit estimators is that in our pairwise comparison setting, each data tuple is used twice – as $(X,Y^{(1)},Y^{(2)},Z)$ and $(X,Y^{(2)},Y^{(1)},1-Z)$ – in constructing the estimating function. This effectively reduces the variance of the resulting estimator. As a result, we will formally show in Section 5 that our DR estimator is semiparametrically efficient. Additionally, we will establish the consistency of (9) when either $\widehat{g}$ or $\widehat{\pi}_{\text{ref}}$ is correctly specified.

Preference optimization. For preference optimization, our goal is to identify the optimal policy that maximizes the average total preference $p^*(\pi)$. Under the BT model assumption, it is immediate to see that the argmax is equivalent to π^* defined in (2). Given the proposed DR estimator, we estimate the optimal policy by solving

$$\begin{aligned}\widehat{\pi} = \arg\max_{\pi\in\Pi}\Big\{&\widehat{p}_{\text{DR}}(\pi) \\ &-\beta\mathbb{E}_{X\sim\mathcal{D}}D_{\text{KL}}[\pi(\bullet\mid X)\,\|\,\widehat{\pi}_{\text{ref}}(\bullet\mid X)]\Big\}.\end{aligned} \qquad (10)$$

We refer to (10) as DRPO, short for doubly robust preference optimization. Theoretically, we will show in Section 5 that our estimated policy $\widehat{\pi}$ achieves a smaller regret bound than PPO- and DPO-type algorithms when the BT assumption holds. Practically, we implement three refinements to stabilize the training: (i) clipping the IS ratio to avoid extremely large IS ratio; (ii) designing a pseudo objective function to enable Monte Carlo sampling from the target policy during optimization; (iii) adopting the KL divergence measure from the group relative policy optimization (Shao et al., 2024) for variance reduction. Details are relegated to Appendix B to save space.

5. Theoretical Analysis

We begin with a summary of our theories; Figure **??** outlines the roadmap. Our theories are concerned with (i) the MSE of our preference evaluation estimator $\widehat{p}_{\text{DR}}(\pi)$ (see (9)), and (ii) the regret bounds of $\widehat{\pi}$ (see (10)) computed by the proposed preference optimization algorithm. Specifically, Theorem 2 provides a finite sample upper bound for the MSE of $\widehat{p}_{\text{DR}}(\pi)$, which in turn yields its double robustness (Corollary 3) and semi-parametric efficiency (Corollary 4). Meanwhile, Theorem 5 upper bounds the difference in total preference between the optimal in-class policy and $\widehat{\pi}$, without assuming the BT model holds. It reveals the double robustness property of our preference optimization algorithm (Corollary 6). When the BT model holds, Theorem 7 further upper bounds the suboptimal gap of $\widehat{\pi}$, demonstrating that it general achieves smaller gaps than PPO- and DPO-based algorithms.

We next introduce some technical conditions.

Assumption 1 (Coverage)**.** Both π/π_{ref} and $\pi/\widehat{\pi}_{\text{ref}}$ are upper bounded by ϵ^{-1} for some constant $\epsilon > 0$.

Assumption 2 (Boundedness)**.** When the BT model holds, both the oracle reward function r^* and its estimator are bounded functions.

Assumption 3 (Realizability)**.** When the BT model holds, π^* that maximizes the expected reward (see (2)) belongs to the parameterized policy class Π in (10).

Assumption 4 (Model complexity)**.** Π belongs to the VapnikChervonenkis (VC) type class (Chernozhukov et al., 2014, Definition 2.1) with a finite VC index $v > 0$.

We remark that similar coverage, boundedness and realizability assumptions are commonly imposed in the OPE and RL literature (see e.g., Chen & Jiang, 2019; Fan et al., 2020; Uehara et al., 2022). The VC-class condition is also frequently assumed in statistics and machine learning (see e.g., Van Der Vaart et al., 1996; Shalev-Shwartz & Ben-David, 2014).

MSE of $\widehat{p}_{\text{DR}}(\pi)$. We next study the statistical properties of the proposed preference estimator $\widehat{p}_{\text{DR}}(\pi)$. Without loss of generality, we also assume both $\widehat{\pi}_{\text{ref}}$ and $\widehat{g}$ (or $\widehat{r}$, in the case where the BT model holds) are obtained from external models independent of $\mathcal{D}$. This condition is mild. Even when such external models are not available and $\widehat{\pi}_{\text{ref}}$ and $\widehat{g}$ are learned internally from $\mathcal{D}$, independence can be preserved using sample-splitting and cross-fitting (Chernozhukov et al., 2018).

Theorem 2 (MSE)**.** *Under Assumption 1, with n data tuples, the semi-parametric efficiency bound (SEB) for estimating $p^*(\pi)$ is given by $n^{-1}Var(\psi(X, Y^{(1)}, Y^{(2)}, Z; \pi, \pi_{ref}, g^*))$. Additionally, the MSE of our $\widehat{p}_{DR}(\pi)$ equals*

$$SEB + O\left(\frac{1}{n}\|\widehat{g} - g^*\|\right) + O\left(\frac{1}{n}\|\frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}} - 1\|\right) + O\left(\|\frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}} - 1\|^2 \cdot \|\widehat{g} - g^*\|^2\right), \qquad (11)$$

where $\|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\|$ and $\|\widehat{g} - g^*\|$ denote the root mean squared errors of $\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}}$ and $\widehat{g}$; see Appendix A.3 for their definitions.

The first part of Theorem 2 establishes the SEB – the smallest-possible MSE that one can hope for estimating $p^*(\pi)$. The second part upper bounds the excess MSE of our estimator over SEB. Specifically, this excess MSE consists of three parts: the first two are excess variance terms arising from estimation errors in the reference policy and the preference model, while the third is a bias term introduced by these estimation errors. Notably, both variance terms scales as $O(n^{-1})$, whereas the SEB itself is also of the order $O(n^{-1})$ under the coverage assumption given ϵ is a constant. Consequently, when either $\widehat{\pi}_{\text{ref}}$ or $\widehat{g}$ is correctly specified, the MSE of $\widehat{p}_{\text{DR}}(\pi)$ coverages to zero as the sample n approaches to infinity. This establishes the double robustness property of our estimator, which we state below.

Corollary 3 (Doubly robust evaluation)**.** *Under Assumption 1, when either $\widehat{\pi}_{\text{ref}}$ or $\widehat{g}$ is correctly specified, the MSE of $\widehat{p}_{DR}(\pi)$ decays to zero as n approaches to infinity.*

We next consider the case where both $\widehat{\pi}_{\text{ref}}$ and $\widehat{g}$ are "approximately" correct in that both root MSEs $\|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\|$ and $\|\widehat{g} - g^*\|$ decay to zero as $n \to \infty$. Since SEB is of the order $O(n^{-1})$, the first two variance terms in (11) decay to zero at a much faster rate than SEB. Meanwhile, when the product $\|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\|\|\widehat{g} - g^*\| = o(n^{-1/2})$, the last bias term in (11) becomes negligible compared to SEB as well. Together, these conditions imply that the MSE of $\widehat{p}_{\text{DR}}(\pi)$ asymptotically matches the SEB, which establishes the semi-parametric efficiency of our estimator. We also remark that conditions similar to $\|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\|\|\widehat{g} - g^*\| = o(n^{-1/2})$ are widely assumed in the literature (see e.g., Chernozhukov et al., 2018; Farrell et al., 2021; Kallus & Uehara, 2022).

Corollary 4 (Semi-parametric efficiency)**.** *Under Assumption 1, when both $\|\frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}} - 1\|$ and $\|\widehat{g} - g^*\|$ decay to zero as $n \to \infty$, and their product is $o(n^{-1/2})$, then MSE($\widehat{p}_{DR}(\pi)$)/SEB $\to 1$ as $n \to \infty$.*

Regret of $\widehat{\pi}$. Next, we derive the statistical properties of the proposed policy $\widehat{\pi}$. When the BT model assumption is violated, we measure the regret of a given policy π using the gap between the total preference of the best in-class policy and that of π, i.e., $\text{Reg}(\pi) = \sup_{\pi' \in \Pi} p^*(\pi') - p^*(\pi)$. By definition, a smaller regret indicates a better policy.

Theorem 5 (Regret)**.** *Under Assumptions 1 (assuming it holds for any $\pi \in \Pi$) and 4, then*

$$Reg(\widehat{\pi}) = O\Big(\beta + \sqrt{\frac{v}{n}} + \frac{v}{n} + \|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\|\|\widehat{g} - g^*\|\Big). \qquad (12)$$

It can be seen from (12) that the regret bound depends on several factors: (i) it decays with the sample size n; (ii) it increases with the regularization parameter β in the KL divergence penalty; (iii) it increases with v, which measures the complexity of the policy class; (iv) it decreases with the estimating error of the reference policy and the preference model. Crucially, the last dependence appears as the product $\|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\|\|\widehat{g} - g^*\|$, which enables us to establish the double robustness property in the context of preference optimization.

Corollary 6 (Doubly robust optimization)**.** *Suppose $\beta \to 0$ as $n \to 0$. Under the conditions in Theorem 5, when either $\widehat{\pi}_{\text{ref}}$ or $\widehat{g}$ is correctly specified, the regret of $\widehat{\pi}$ decays to zero as $n \to \infty$.*

Finally, we restrict our attention to the ideal setting where the BT model holds and upper bound the suboptimality gap, defined as the difference in the expected reward between the optimal policy π^* and our $\widehat{\pi}$, i.e., $J(\pi^*) - J(\widehat{\pi})$.

Theorem 7 (Suboptimality gap)**.** *Suppose the BT model assumption in (1) holds. Under Assumptions 2, 3, and the conditions in Theorem 5, the suboptimality gap of $\widehat{\pi}$ is upper bounded by*

$$O\Big(\beta + \sqrt{\frac{v}{n}} + \frac{v}{n} + \|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\|\|\widehat{r} - r^*\|\Big). \quad (13)$$

Meanwhile, for PPO-based algorithms, their suboptimality gaps are bounded by

$$O\Big(\beta + \sqrt{\frac{v}{n}} + \frac{v}{n} + \|\widehat{r} - r^*\|\Big). \quad (14)$$

Finally, for DPO-based algorithms, their suboptimality gaps are bounded by

$$O\Big(\exp(-\bar{c}\beta^{-1}) + \frac{1}{\beta}\sqrt{\frac{v}{n}} + \|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\|\Big), \quad (15)$$

for some constant $\bar{c} > 0$, under conditions specified in Appendix A.7.

According to (13) and (14) that, by using a sufficiently small β, the suboptimality gaps of PPO-based and our algorithms are of the order $O(n^{-1/2} + \|\widehat{r} - r^*\|)$ and $O(n^{-1/2} + \|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\|\|\widehat{r} - r^*\|)$, respectively. As for DPO-based algorithms, For DPO-based algorithms, setting $\beta = \bar{c}^{-1}C\log n$ for some constant $C > 0$ makes the first term in (15) of order $O(n^{-C})$, which can be made arbitrarily small with a sufficiently large C. The second term remains of order $O(n^{-1/2})$ up to a logarithmic factor, yielding an overall suboptimality gap of $O(n^{-1/2}\log n + \|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\|)$. Consequently, our algorithm's suboptimality gap is more robust to estimation errors in the reference policy and preference model, as these errors influence our bound only through their product. In contrast, for PPO- and DPO-based algorithms, these errors affect their suboptimality bounds in the first order. In particular, when these errors converge to zero at a rate of $O(n^{-c})$ for some $0 < c < 1/2$, our algorithm achieves strictly smaller suboptimality bounds than both DPO- and PPO-based algorithms.

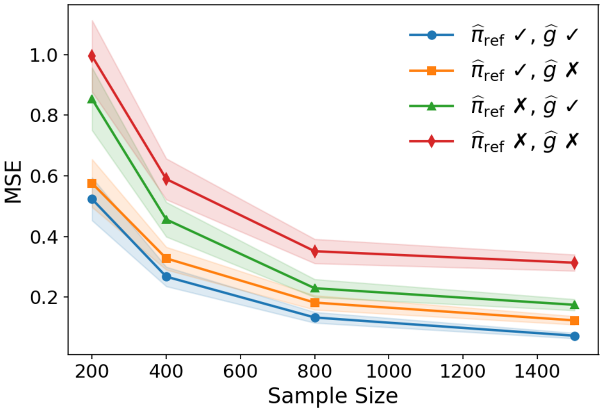


Figure 2. MSEs of different preference evaluation estimators on the IMDb dataset. Shaded areas visualize the 95% confidence bands.

6. Experiments

In this section, we first use the IMDb dataset (Maas et al., 2011) to empirically validate the double robustness property of our preference estimator $\widehat{p}_{\text{DR}}$ (Equation 9) established in Corollary 3. We next compare the proposed preference optimization algorithm (Equation 10) against baseline approaches on the *Too Long; Didn't Read* (TL;DR, Völske et al., 2017) and *Anthropic Helpful and Harmless* (HH, Bai et al., 2022) datasets.

6.1. Preference Evaluation

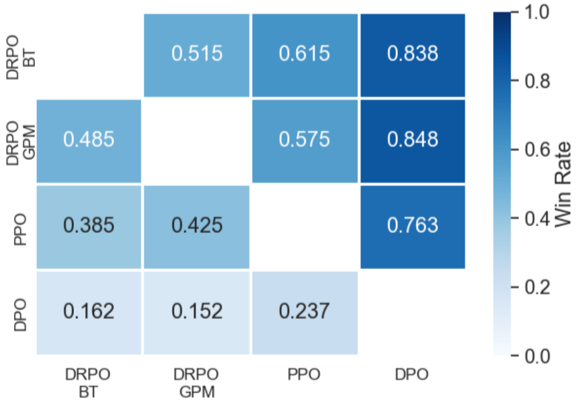
Task and objective. We consider the *controlled sentiment generation* task which aims to produce positive movie reviews using the IMDb dataset. We first apply SFT to the `EleutherAI/gpt-neo-125m` base model (Black et al., 2021), which serves as the reference policy for response generation. The generated responses are then annotated using a pre-trained sentiment classifier to produce preference labels. Using these synthetic data, we train an optimal policy via DPO. Our objective in this section is to evaluate the total preference of this DPO-trained policy over the SFT-based reference policy. Its oracle value, computed via Monte Carlo, is 0.681. Additional details on data generation and model training are provided in Appendix C.1.

Evaluation and result. To empirically assess the double robustness property, we evaluate four variants of our preference estimator, each with either the preference model and/or the reference policy correctly specified or misspecified. To misspecify the preference model, we set $\widehat{g}$ to a uniformly random value in $[0, 1]$. To misspecify the reference policy, we use the unfine-tuned `EleutherAI/gpt-neo-125m` base model. Figure 2 displays the MSEs (solid lines on left panel) and their associated 95% confidence intervals (shaded areas) of the four estimators across different sample sizes, averaged over 500 simulations. It can be seen that the estimator with both models misspecified (red line) exhibits a significantly larger MSE than the other three and shows minimal improvement beyond 800 samples. To the contrary, when either the preference model or the reference policy is correctly specified (yellow and green lines), the MSE is substantially reduced with a moderately large sample size. This aligns with the double robustness property. Meanwhile, the estimator with both correctly specified models (blue line) achieves the lowest MSE (being very close to zero with 1500 data tuples), supporting its semiparametric efficiency.

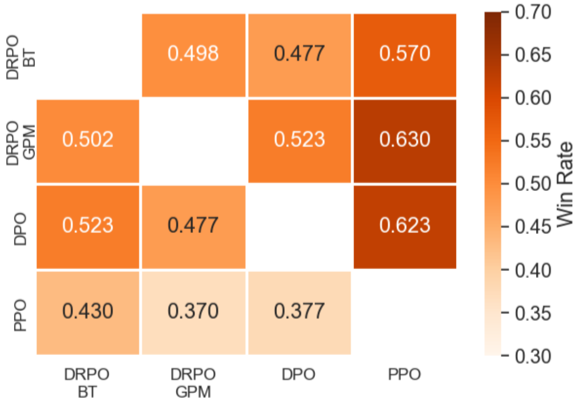
6.2. Preference Optimization

Tasks and objective. This section considers two tasks: *summarization* and *human dialogue*. First, for summarization, we use the TL;DR dataset with preference annotations from (Stiennon et al., 2020) (lengthy Reddit posts as prompts, behavioral SFT models' summaries as responses) to fine-tune LLMs for concise, informative summaries. The SFT and reward models for this task are from `cleanrl` (Huang et al., 2024), with the SFT model trained on a similar but different dataset. Second, for human dialogue, the HH dataset (human queries as prompts) is used to align LLMs for helpful responses. Its SFT and reward models are trained using `TRL` (von Werra et al., 2020) framework due to the unavailability of pre-existing ones. For each task, a reward-based BT preference model (using the same reward model for PPO training) and a general preference model (Zhang et al., 2024d) are adopted to serve as $\widehat{g}$ (donating as DRPO-BT and DRPO-GPM). Refer to more details of the implementation and baseline training in Appendix C.2.

Evaluation and result. Given the absence of ground-truth preference or reward models, we follow prior work (Rafailov et al., 2023; Wu et al., 2024a; Ye et al., 2025) and use `GPT-4o-mini` to assess the quality of generated responses (details in Appendix C.2). Win rates at default temperature 1.0 are presented in Figure 3, with results at other temperatures presented in Appendix D. In *summarization*, both DRPO-BT and DRPO-GPM largely outperform DPO and PPO. For this task, we use an external reference policy model for $\widehat{\pi}_{\text{ref}}$, which is likely misspecified due to being trained on a different dataset. This likely contributes to DPO's poor performance. However, despite relying on the same reference policy, the superior performance of our algorithms highlights their robustness to reference policy misspecification. In *human dialogue*, DRPO-GPM demonstrates the best performance, whereas DRPO-BT outperforms PPO and achieves comparable performance to DPO. The relatively weaker performance of PPO suggests potential misspecification in the reward model for this task. Despite using the same reward model, DRPO-BT effectively mitigates this challenge and achieves a win rate of 57% against PPO, demonstrating its robustness. Finally, it is worth highlighting that the hyperparameters for our method are not extensively tuned, unlike the baselines – particularly PPO, for which the best-performing version is reported – further demonstrating the robustness of our proposal.



(a) TLDR



(b) HH

Figure 3. **Pairwise win rate** matrices between different methods across two datasets. Each entry indicates how often the row method outperforms the column method; higher values denote better performance.

References

Agranov, M. and Ortoleva, P. Stochastic choice and preferences for randomization. *Journal of Political Economy*, 125:40 – 68, 2015. URL `https://api.semanticscholar.org/CorpusID:11440933`.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

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 *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, 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.

Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Bibaut, A., Malenica, I., Vlassis, N., and Van Der Laan, M. More efficient off-policy evaluation through regularized targeted learning. In *International Conference on Machine Learning*, pp. 654–663. PMLR, 2019.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York, 1998. ISBN 978-0-387-98473-5. URL `https://link.springer.com/book/9780387984735`.

Black, S., Leo, G., Wang, P., Leahy, C., and Biderman, S. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL `https://doi.org/10.5281/zenodo.5297715`. If you use this software, please cite it using these metadata.

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

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

Calandriello, D., Guo, D., Munos, R., Rowland, M., Tang, Y., Pires, B. A., Richemond, P. H., Lan, C. L., Valko, M., Liu, T., et al. Human alignment of large language models through online preference optimisation. *arXiv preprint arXiv:2403.08635*, 2024.

Cao, D. and Zhou, A. Orthogonalized estimation of difference of q-functions. *arXiv preprint arXiv:2406.08697*, 2024.

Chan, A. J., Sun, H., Holt, S., and Van Der Schaar, M. Dense reward for free in reinforcement learning from human feedback. *arXiv preprint arXiv:2402.00782*, 2024.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Chen, Y., Zhu, D., Sun, Y., Chen, X., Zhang, W., and Shen, X. The accuracy paradox in rlhf: When better reward models don't yield better language models. *arXiv preprint arXiv:2410.06554*, 2024.

Chernozhukov, V., Chetverikov, D., and Kato, K. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, pp. 1564–1597, 2014.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

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

Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S., and Zhu, R. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):179–211, 2023.

Cui, Y., Pu, H., Shi, X., Miao, W., and Tchetgen Tchetgen, E. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359, 2024.

Dudík, M., Erhan, D., Langford, J., and Li, L. Doubly Robust Policy Evaluation and Optimization. *Statistical Science*, 29(4):485 – 511, 2014. doi: 10.1214/14-STS500.

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., and Madry, A. Implementation matters in deep rl: A case study on ppo and trpo. In *International conference on learning representations*, 2019.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Fan, C., Lu, W., Song, R., and Zhou, Y. Concordance-assisted learning for estimating optimal individualized treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(5):1565–1582, 2017.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. A theoretical analysis of deep q-learning. In *Learning for dynamics and control*, pp. 486–489. PMLR, 2020.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. *International Conference on Machine Learning*, pp. 1447–1456, 2018.

Farrell, M. H., Liang, T., and Misra, S. Deep neural networks for estimation and inference. *Econometrica*, 89 (1):181–213, 2021.

Fu, J., Zhao, X., Yao, C., Wang, H., Han, Q., and Xiao, Y. Reward shaping to mitigate reward hacking in rlhf. *arXiv preprint arXiv:2502.18770*, 2025.

Fulcher, I. R., Shpitser, I., Marealle, S., and Tchetgen Tchetgen, E. J. Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):199–214, 2020.

Gao, J., Xu, S., Ye, W., Liu, W., He, C., Fu, W., Mei, Z., Wang, G., and Wu, Y. On designing effective rl reward at training time for llm reasoning. *arXiv preprint arXiv:2410.15115*, 2024.

Gardner, M. *Mathematical games, the paradox of the nontransitive dice and the elusive principle of indifference*. Dec 1970. URL https://www.scientificamerican.com/article/mathematical-games-1970-12/.

Gorbatovski, A., Shaposhnikov, B., Malakhov, A., Surnachev, N., Aksenov, Y., Maksimov, I., Balagansky, N., and Gavrilov, D. Learn your reference model for real good alignment. *arXiv preprint arXiv:2404.09656*, 2024.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Han, L., Shen, Z., and Zubizarreta, J. Multiply robust federated estimation of targeted average treatment effects. *Advances in Neural Information Processing Systems*, 36: 70453–70482, 2023.

Hartmann, J., Heitmann, M., Siebert, C., and Schamp, C. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87, 2023. doi: https://doi.org/10.1016/j.ijresmar.2022.05.005. URL https://www.sciencedirect.com/science/article/pii/S0167811622000477.

Hong, Y., Zhang, H., Bao, J., Jiang, H., and Song, Y. Energy-based preference model offers better offline alignment than the bradley-terry preference model. *arXiv preprint arXiv:2412.13862*, 2024.

Hu, J. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.

Huang, S., Noukhovitch, M., Hosseini, A., Rasul, K., Wang, W., and Tunstall, L. The n+ implementation details of rlhf with ppo: A case study on tl; dr summarization. *arXiv preprint arXiv:2403.17031*, 2024.

Imai, K. and Ratkovic, M. Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):243–263, 2014.

Jiang, N. and Li, L. Doubly robust off-policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Jiang, R., Lu, W., Song, R., and Davidian, M. On estimation of optimal treatment regimes for maximizing t-year survival probability. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1165–1185, 2017.

Kallus, N. and Uehara, M. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21: 1–63, 2020a.

Kallus, N. and Uehara, M. Statistically efficient off-policy policy gradients. In *International Conference on Machine Learning*, pp. 5089–5100. PMLR, 2020b.

Kallus, N. and Uehara, M. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Oper. Res.*, 70(6):32823302, November 2022. ISSN 0030-364X.

Kallus, N. and Zhou, A. Policy evaluation and optimization with continuous treatments. In *International conference on artificial intelligence and statistics*, pp. 1243–1251. PMLR, 2018.

Kaufmann, T., Weng, P., Bengs, V., and Hüllermeier, E. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 10, 2023.

Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4): 1229–1245, 2017.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Laidlaw, C., Singhal, S., and Dragan, A. Correlated proxies: A new definition and improved mitigation for reward hacking. *arXiv preprint arXiv:2403.03185*, 2024.

Li, H., Zheng, C., and Wu, P. Stabledr: Stabilized doubly robust learning for recommendation on data missing not at random. *arXiv preprint arXiv:2205.04701*, 2022.

Li, X., Miao, W., Lu, F., and Zhou, X.-H. Improving efficiency of inference in clinical trials with external control data. *Biometrics*, 79(1):394–403, 2023a.

Li, Z., Xu, T., Zhang, Y., Lin, Z., Yu, Y., Sun, R., and Luo, Z.-Q. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. *arXiv preprint arXiv:2310.10505*, 2023b.

Liao, P., Qi, Z., Wan, R., Klasnja, P., and Murphy, S. A. Batch policy learning in average reward markov decision processes. *Annals of statistics*, 50(6):3364, 2022.

Lindner, D. and El-Assady, M. Humans are not boltzmann distributions: Challenges and opportunities for modelling human feedback and interaction in reinforcement learning. *arXiv preprint arXiv:2206.13316*, 2022.

Liu, K., Long, Q., Shi, Z., Su, W. J., and Xiao, J. Statistical impossibility and possibility of aligning llms with human preferences: From condorcet paradox to nash equilibrium. *arXiv preprint arXiv:2503.10990*, 2025a.

Liu, P., Shi, C., and Sun, W. W. Dual active learning for reinforcement learning from human feedback. *arXiv preprint arXiv:2410.02504*, 2024a.

Liu, Y., Liu, P., and Cohan, A. Understanding reference policies in direct preference optimization. *arXiv preprint arXiv:2407.13709*, 2024b.

Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S., and Lin, M. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025b.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019. URL `https://arxiv.org/abs/1711.05101`.

Luedtke, A. R. and Van Der Laan, M. J. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2): 713, 2016.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1015`.

May, K. O. Intransitivity, utility, and the aggregation of preference patterns. *Econometrica*, 22:1, 1954. URL `https://api.semanticscholar.org/CorpusID:156169619`.

Michaud, E. J., Gleave, A., and Russell, S. Understanding learned reward functions. *arXiv preprint arXiv:2012.05862*, 2020.

Milano, S., Taddeo, M., and Floridi, L. Ethical aspects of multi-stakeholder recommendation systems. *The information society*, 37(1):35–45, 2021.

Munos, R., Valko, M., Calandriello, D., Azar, M. G., Rowland, M., Guo, Z. D., Tang, Y., Geist, M., Mesnard, T., Michi, A., et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 18, 2023.

Newey, W. K. Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135, 1990.

Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Oprescu, M., Syrgkanis, V., and Wu, Z. S. Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, pp. 4932–4941. PMLR, 2019.

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

Quinzan, F., Soleymani, A., Jaillet, P., Rojas, C. R., and Bauer, S. Drcfs: Doubly robust causal feature selection. In *International Conference on Machine Learning*, pp. 28468–28491. PMLR, 2023.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741, 2023.

Robins, J. M. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics: analysis of correlated data*, pp. 189–326. Springer, 2004.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

Robins, J. M., Li, L., Mukherjee, R., Tchetgen, E. T., and van der Vaart, A. Minimax estimation of a functional on a structured high-dimensional model. *THE ANNALS of STATISTICS*, pp. 1951–1987, 2017.

Rosset, C., Cheng, C.-A., Mitra, A., Santacroce, M., Awadallah, A., and Xie, T. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4): 640, 2015.

Shah, R. D. and Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Shi, C., Fan, A., Song, R., and Lu, W. High-dimensional a-learning for optimal dynamic treatment regimes. *Annals of statistics*, 46(3):925, 2018.

Shi, C., Blei, D., and Veitch, V. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Shi, C., Xu, T., Bergsma, W., and Li, L. Double generative adversarial networks for conditional independence testing. *Journal of Machine Learning Research*, 22(285): 1–32, 2021.

Shi, C., Uehara, M., Huang, J., and Jiang, N. A minimax learning approach to off-policy evaluation in confounded partially observable markov decision processes. In *International Conference on Machine Learning*, pp. 20057–20094. PMLR, 2022.

Shi, C., Luo, S., Le, Y., Zhu, H., and Song, R. Statistically efficient advantage learning for offline reinforcement learning in infinite horizons. *Journal of the American Statistical Association*, 119(545):232–245, 2024a.

Shi, C., Zhu, J., Shen, Y., Luo, S., Zhu, H., and Song, R. Off-policy confidence interval estimation with confounded markov decision process. *Journal of the American Statistical Association*, 119(545):273–284, 2024b.

Skalse, J., Howe, N., Krasheninnikov, D., and Krueger, D. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

Song, F., Yu, B., Li, M., Yu, H., Huang, F., Li, Y., and Wang, H. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18990–18998, 2024.

Song, R., Luo, S., Zeng, D., Zhang, H. H., Lu, W., and Li, Z. Semiparametric single-index model for estimating optimal individualized treatment strategy. *Electronic journal of statistics*, 11(1):364, 2017.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 3008–3021, 2020.

Su, Y., Dimakopoulou, M., Krishnamurthy, A., and Dudík, M. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pp. 9167–9176. PMLR, 2020.

Sutton, R. S., Barto, A. G., et al. *Reinforcement learning: An introduction*. MIT press Cambridge, 2018.

Swamy, G., Dann, C., Kidambi, R., Wu, Z. S., and Agarwal, A. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

Tan, Z. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, 2010.

Tang, Y., Guo, Z. D., Zheng, Z., Calandriello, D., Munos, R., Rowland, M., Richemond, P. H., Valko, M., Pires, B. Á., and Piot, B. Generalized preference optimization: A unified approach to offline alignment. *arXiv preprint arXiv:2402.05749*, 2024.

Tang, Z., Feng, Y., Li, L., Zhou, D., and Liu, Q. Doubly robust bias reduction in infinite horizon off-policy estimation. In *International Conference on Learning Representations*, 2020.

Team, Q. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Thomas, P. S. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148. PMLR, 2016.

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

Tsiatis, A. A. *Semiparametric Theory and Missing Data*. Springer, 2006.

Tsiatis, A. A., Davidian, M., and Cao, W. Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. *Biometrics*, 67(2):536–545, 2011.

Tversky, A. Intransitivity of preferences. *Psychological Review*, 76:31–48, 1969. URL https://api.semanticscholar.org/CorpusID:144609998.

Uehara, M., Huang, J., and Jiang, N. Minimax weight and q-function learning for off-policy evaluation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9659–9668. PMLR, 13–18 Jul 2020a.

Uehara, M., Kato, M., and Yasui, S. Off-policy evaluation and learning for external validity under a covariate shift. *Advances in Neural Information Processing Systems*, 33: 49–61, 2020b.

Uehara, M., Shi, C., and Kallus, N. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

Van Der Vaart, A. W., Wellner, J. A., van der Vaart, A. W., and Wellner, J. A. *Weak convergence*. Springer, 1996.

Vermeulen, K. and Vansteelandt, S. Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036, 2015.

Völske, M., Peyrard, M., Bevendorff, J., Potthast, M., and Stein, B. TL;DR: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization (EMNLP)*, pp. 59–63. Association for Computational Linguistics, 2017.

von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., Huang, S., Rasul, K., and Gallouédec, Q. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

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

Wang, J., Wang, H., Sun, S., and Li, W. Aligning language models with human preferences via a bayesian approach. *Advances in Neural Information Processing Systems*, 36: 49113–49132, 2023b.

Wang, L. and Han, P. Multiply robust estimation for average treatment effect among treated. *Statistical Theory and Related Fields*, 8(1):29–39, 2024.

Wang, L. and Tchetgen Tchetgen, E. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80 (3):531–550, 2018.

Wang, Y. and Shah, R. D. Debiased inverse propensity score weighting for estimation of average treatment effects with high-dimensional confounders. *The Annals of Statistics*, 52(5):1978–2003, 2024.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp.

38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Wu, T., Zhu, B., Zhang, R., Wen, Z., Ramchandran, K., and Jiao, J. Pairwise proximal policy optimization: Language model alignment with comparative rl. In *First Conference on Language Modeling*, 2024a.

Wu, Y., Sun, Z., Yuan, H., Ji, K., Yang, Y., and Gu, Q. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*, 2024b.

Xiao, J., Li, Z., Xie, X., Getzen, E., Fang, C., Long, Q., and Su, W. J. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization. *arXiv preprint arXiv:2405.16455*, 2024.

Xiao, T., Yuan, Y., Li, M., Chen, Z., and Honavar, V. G. On a connection between imitation learning and rlhf. *arXiv preprint arXiv:2503.05079*, 2025.

Xie, C., Yang, W., and Zhang, Z. Semiparametrically efficient off-policy evaluation in linear markov decision processes. In *International Conference on Machine Learning*, pp. 38227–38257. PMLR, 2023.

Xu, W., Li, J., Wang, W. Y., and Li, L. Bpo: Staying close to the behavior llm creates better online llm alignment. *arXiv preprint arXiv:2406.12168*, 2024.

Xu, Y., Zhu, J., Shi, C., Luo, S., and Song, R. An instrumental variable approach to confounded off-policy evaluation. In *International Conference on Machine Learning*, pp. 38848–38880. PMLR, 2023.

Ye, C., Xiong, W., Zhang, Y., Dong, H., Jiang, N., and Zhang, T. Online iterative reinforcement learning from human feedback with general preference model. *Advances in Neural Information Processing Systems*, 37: 81773–81807, 2024.

Ye, K., Zhou, H., Zhu, J., Quinzan, F., and Shi, C. Robust reinforcement learning from human feedback for large language models fine-tuning. *arXiv preprint arXiv:2504.03784*, 2025.

Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Fan, T., Liu, G., Liu, L., Liu, X., et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yuan, Y., Yu, Q., Zuo, X., Zhu, R., Xu, W., Chen, J., Wang, C., Fan, T., Du, Z., Wei, X., et al. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3): 681–694, 2013.

Zhang, H., Zhou, S., Guan, J., and Huan, J. Measuring conditional independence by independent residuals for causal discovery. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–19, 2019.

Zhang, H., Lei, Y., Gui, L., Yang, M., He, Y., Wang, H., and Xu, R. Cppo: Continual learning for reinforcement learning with human feedback. In *The Twelfth International Conference on Learning Representations*, 2024a.

Zhang, Y., Huang, L., Yang, Y., and Shao, X. Doubly robust conditional independence testing with generative neural networks. *arXiv preprint arXiv:2407.17694*, 2024b.

Zhang, Y., Yu, D., Peng, B., Song, L., Tian, Y., Huo, M., Jiang, N., Mi, H., and Yu, D. Iterative nash policy optimization: Aligning llms with general preferences via no-regret learning. *arXiv preprint arXiv:2407.00617*, 2024c.

Zhang, Y., Zhang, G., Wu, Y., Xu, K., and Gu, Q. General preference modeling with preference representations for aligning language models. *arXiv preprint arXiv:2410.02197*, 2024d.

Zhao, Y., Khalman, M., Joshi, R., Narayan, S., Saleh, M., and Liu, P. J. Calibrating sequence likelihood improves conditional language generation. *arXiv preprint arXiv:2210.00045*, 2022.

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

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A. Technical Proof

In this section, we present the regularity conditions and proofs for all the lemmas and theorems. By nature, the vocabulary size is finite; as such, all random variables – including the prompts X and the responses Y – are discrete. We assume that ϵ in the coverage assumption is a constant , which is why it does not explicitly appear in the error bound. However, in the proof of Theorems, for completeness, we will explicitly highlight how the leading terms of the error bounds depend on ϵ.

A.1. Proof of Lemma 1

By direct calculation, it follows that

$$\begin{aligned}\mathbb{E}\left\{w(Y^{(1)},X)Z\right\} &= \mathbb{E}\left\{\mathbb{E}\left[\frac{\pi(Y^{(1)}|X)}{\pi_{\text{ref}}(Y^{(1)}|X)}\mathbb{I}\{Y^{(1)}\succ Y^{(2)}\}\middle|X,Y^{(1)},Y^{(2)}\right]\right\}\\ &= \mathbb{E}\left\{\frac{\pi(Y^{(1)}|X)}{\pi_{\text{ref}}(Y^{(1)}|X)}g^*\left(Y^{(1)},Y^{(2)},X\right)\right\}\\ &= \mathbb{E}\left\{\sum_y \pi(y|X)g^*\left(y,Y^{(2)},X\right)\right\}\\ &= \mathbb{E}\left\{\mathbb{E}_{y\sim\pi(\bullet|X)}g^*\left(y,Y^{(2)},X\right)\right\},\end{aligned}$$

where the first equality is derived by the law of total expectation, the second equality follows from the definition of the preference function g^*, and the third equality follows from the change-of-measure theorem (e.g., RadonNikodym theorem).

Following a similar argument and using the fact that $1-Z=\mathbb{I}(Y^{(2)}\succ Y^{(1)})$, we obtain

$$\mathbb{E}\left\{w(Y^{(2)},X)(1-Z)\right\} = \mathbb{E}\left\{\mathbb{E}_{y\sim\pi(\bullet|X)}g^*\left(y,Y^{(1)},X\right)\right\}.$$

Consequently, $p^*(\pi)=\frac{1}{2}\mathbb{E}[w(Y^{(1)},X)Z+w(Y^{(2)},X)(1-Z)]$, which finishes the proof of the lemma.

A.2. Auxiliary lemma for proving Theorem 2

Before proceeding to the proof of Theorem 2, we first introduce an auxiliary lemma.

Lemma A8. *Under Assumption 1, with n independent data tuple $W_i=(X_i,Y_i^{(1)},Y_i^{(2)},Z_i), i=1,\dots n$, the efficient influence function (see e.g., Tsiatis, 2006, for the detailed definition) for $p^*(\pi)$ is given by $\frac{1}{n}\sum_{i=1}^n\psi(X_i,Y_i^{(1)},Y_i^{(2)},Z_i;\pi,\pi_{\text{ref}},g^*)-p^*(\pi)$, with ψ defined in equation (8).*

Proof of Lemma A8. To simplify notation, we denote $\psi(W)=\psi(X,Y^{(1)},Y^{(2)},Z;\pi,\pi_{\text{ref}},g^*)$. Let $\mathcal{M}$ denote the model that generates these data triplets, which are i.i.d. copies of $W=(Z,Y^{(1)},Y^{(2)},X)$. This model involves three types of parameters: (i) those to model the probability mass function $f_X(\bullet)$ of the prompt X (denoted by γ); (ii) those to model the reference policy which generates response $Y^{(1)},Y^{(2)}$ independently conditional on the prompt X (denoted by b) and (iii) those to model the preference probability g^* which characterize the probability of $Y^{(1)}$ is preferred than $Y^{(2)}$ given X (denoted by η). Then the likelihood function for a data tuple W is given by

$$l(W;\gamma,b,\eta)=f_\gamma(X)\pi_b(Y^{(1)}|X)\pi_b(Y^{(2)}|X)g_\eta(Y^{(1)},Y^{(2)},X)^Z(1-g_\eta(Y^{(1)},Y^{(2)},X))^{1-Z}. \tag{16}$$

Additionally, let (γ_0,b_0,η_0) denote the true parameters in the model so that $f_{\gamma_0}=f_X,\pi_{b_0}=\pi_{\text{ref}}$ and $g_{\eta_0}=g^*$.

The proof follows from standard techniques in semi-parametric statistic; see e.g., Chapters 2 & 3 in Bickel et al. (1998) and Theorem 3.5 in Tsiatis (2006). See also the proof of Theorem 1 in (Kallus & Uehara, 2020a). Specifically:

1. For any given policy π, we first prove that $\mathbb{E}[\{\psi(W)-p^*(\pi)\}\nabla\log l(W;\gamma_0,b_0,\eta_0)]$ is a valid derivative of $p^*(\pi)$ with respect to the parameters (γ_0,b_0,η_0), where ∇ denotes the gradient operator.
2. We next prove that $\psi(W)-p^*(\pi)$ lies in the tangent space of the data generating process model $\mathcal{M}$ (denoted by $\mathcal{T}_{\mathcal{M}}$), that is, $\psi(W)-p^*(\pi)\in\mathcal{T}_{\mathcal{M}}$.

Step 1: $\mathbb{E}[\{\psi(W)-p^*(\pi)\}\nabla\log l(W;\gamma_0,b_0,\eta_0)]$ **is a valid derivative of** $p^*(\pi)$ **with respect to** (γ_0,b_0,η_0).
Noted that the log-likelihood has zero mean. Therefore, in order to prove step 1, we only need to verify the following three equations hold.

$$
\begin{aligned}
&\text{(i)}\ \mathbb{E}\left\{\psi(W)\tfrac{\partial}{\partial\gamma}\log l(W;\gamma_0,b_0,\eta_0)\right\}=\tfrac{\partial}{\partial\gamma}p^*(\pi)|_{\gamma=\gamma_0},\\
&\text{(ii)}\ \mathbb{E}\left\{\psi(W)\tfrac{\partial}{\partial b}\log l(W;\gamma_0,b_0,\eta_0)\right\}=\tfrac{\partial}{\partial b}p^*(\pi)|_{b=b_0},\\
&\text{(iii)}\ \mathbb{E}\left\{\psi(W)\tfrac{\partial}{\partial\eta}\log l(W;\gamma_0,b_0,\eta_0)\right\}=\tfrac{\partial}{\partial\eta}p^*(\pi)_{\eta=\eta_0}.
\end{aligned}
$$

By definition, $p^*(\pi)$ can be represented as

$$
\begin{aligned}
p^*(\pi) &= \mathbb{E}[\mathbb{E}_{y_1\sim\pi_\theta,y_2\sim\pi_{\text{ref}}}\mathbb{P}(y_1\succ y_2|X)]\\
&= \sum_{x,y_1,y_2} g^*(y_1,y_2,x)\pi(y_1|x)\pi_{\text{ref}}(y_2|x)f_X(x).
\end{aligned}
$$

Let $w=(x,y_1,y_2,z)$ denote the realization of $W=(X,Y^{(1)},Y^{(2)},Z)$. It follows from equation (16) that

$$
\begin{aligned}
\log l(w;\gamma,b,\eta) &= \log f_\gamma(x)+\log\pi_b(y_1|x)+\log\pi_b(y_2|x)\\
&\quad +z\log g_\eta(y_1,y_2,x)+(1-z)\log(1-g_\eta(y_1,y_2,x)).
\end{aligned}
\tag{17}
$$

With some calculations, we obtain

$$
\begin{aligned}
\frac{\partial}{\partial\gamma}\log l(w;\gamma_0,b_0,\eta_0) &= \frac{1}{f_X(x)}\frac{\partial}{\partial\gamma}f_\gamma(x)\Big|_{\gamma=\gamma_0},\\
\frac{\partial}{\partial b}\log l(w;\gamma_0,b_0,\eta_0) &= \frac{1}{\pi_{\text{ref}}(y_1|x)}\frac{\partial}{\partial b}\pi_b(y_1|x)\Big|_{b=b_0}+\frac{1}{\pi_{\text{ref}}(y_2|x)}\frac{\partial}{\partial b}\pi_b(y_2|x)\Big|_{b=b_0},\\
\frac{\partial}{\partial\eta}\log l(w;\gamma_0,b_0,\eta_0) &= \left(\frac{z}{g^*(y_1,y_2,x)}-\frac{1-z}{1-g^*(y_1,y_2,x)}\right)\frac{\partial}{\partial\eta}g_\eta(y_1,y_2,x)\Big|_{\eta=\eta_0}.
\end{aligned}
$$

In the following proof, we omit $|_{\gamma=\gamma_0}$, $|_{b=b_0}$ and $|_{\eta=\eta_0}$ to ease notation.

For equation (i): Let Ber(p) denote the Bernoulli distribution with success probability p. The left-hand-side (LHS) of equation (i) can be represented by

$$
\begin{aligned}
&\mathbb{E}\left\{\psi(W)\frac{\partial}{\partial\gamma}\log l(W;\gamma_0,b_0,\eta_0)\right\}\\
=&\ \frac{1}{2}\sum_{x,y_1,y_2}\mathbb{E}_{z\sim\text{Ber}(g^*(y_1,y_2,x))}\left\{\left(\frac{\pi(y_1|x)}{\pi_{\text{ref}}(y_1|x)}-\frac{\pi(y_2|x)}{\pi_{\text{ref}}(y_2|x)}\right)(z-g^*(y_1,y_2,x))\right.\\
&\qquad\left.\times\pi_{\text{ref}}(y_1|x)\pi_{\text{ref}}(y_2|x)\frac{\partial}{\partial\gamma}f_\gamma(x)\right\}\\
&\quad+\frac{1}{2}\sum_{x,y_1,y_2,y^*}\left(g^*(y^*,y_1,x)+g^*(y^*,y_2,x)\right)\pi(y^*|x)\pi_{\text{ref}}(y_1|x)\pi_{\text{ref}}(y_2|x)\frac{\partial}{\partial\gamma}f_\gamma(x)
\end{aligned}
$$

Using the fact that $\mathbb{E}_{z\sim\text{Ber}(g^*(y_1,y_2,x))}\{z-g^*(y_1,y_2,x)\}=0$, the first term on the right-hand-side (RHS) of the above equation vanishes. Therefore,

$$
\begin{aligned}
\mathbb{E}\left\{\psi(W)\frac{\partial}{\partial\gamma}\log l(W;\gamma_0,b_0,\eta_0)\right\} =&\frac{1}{2}\sum_{x,y_1,y^*}g^*(y^*,y_1,x)\pi(y^*|x)\pi_{\text{ref}}(y_1|x)\frac{\partial}{\partial\gamma}f_{\gamma_0}(x)\\
&+\frac{1}{2}\sum_{x,y_2,y^*}g^*(y^*,y_2,x)\pi(y^*|x)\pi_{\text{ref}}(y_2|x)\frac{\partial}{\partial\gamma}f_{\gamma_0}(x)\\
=&\sum_{x,y,y^*}g^*(y^*,y,x)\pi(y^*|x)\pi_{\text{ref}}(y|x)\frac{\partial}{\partial\gamma}f_{\gamma_0}(x)\\
=&\frac{\partial}{\partial\gamma}p^*(\pi).
\end{aligned}
$$

For equation (ii): Notice that the LHS of equation (ii) can be represented as

$$
\begin{aligned}
&\mathbb{E}\left\{\psi(W)\frac{\partial}{\partial b}\log l(W;\gamma_0,b_0,\eta_0)\right\}\\
=\;&\frac{1}{2}\sum_{x,y_1,y_2}\mathbb{E}_{z\sim\text{Ber}(g^*(y_1,y_2,x))}\Bigg\{\left(\frac{\pi(y_1|x)}{\pi_{\text{ref}}(y_1|x)}-\frac{\pi(y_2|x)}{\pi_{\text{ref}}(y_2|x)}\right)\Bigg(\frac{1}{\pi_{\text{ref}}(y_1|x)}\frac{\partial}{\partial b}\pi_b(y_1|x)+\\
&\qquad\frac{1}{\pi_{\text{ref}}(y_2|x)}\frac{\partial}{\partial b}\pi_b(y_2|x)\Bigg)\times(z-g^*(y_1,y_2,x))\pi_{\text{ref}}(y_1|x)\pi_{\text{ref}}(y_2|x)f_X(x)\Bigg\}\\
&+\frac{1}{2}\sum_{x,y_1,y_2,y^*}\left(g^*(y^*,y_1,x)+g^*(y^*,y_2,x)\right)\pi(y^*|x)\frac{\partial}{\partial b}[\pi_{b_0}(y_1|x)\pi_{b_0}(y_2|x)]f_X(x).
\end{aligned}
$$

Follows a similar argument in proving equation (i), the first term on the RHS equals zero. The second term can be further represented by

$$
\begin{aligned}
&\frac{1}{2}\frac{\partial}{\partial b}\sum_{x,y_1,y_2,y^*}\left(g^*(y^*,y_1,x)+g^*(y^*,y_2,x)\right)\pi(y^*|x)\pi_{b_0}(y_1|x)\pi_{b_0}(y_2|x)f_X(x)\\
=\;&\frac{1}{2}\frac{\partial}{\partial b}\sum_{x,y_1,y^*}g^*(y^*,y_1,x)\pi(y^*|x)\pi_{b_0}(y_1|x)f_X(x)\\
&\quad+\frac{1}{2}\frac{\partial}{\partial b}\sum_{x,y^*,y_2}g^*(y^*,y_2,x)\pi(y^*|x)\pi_{b_0}(y_2|x)f_X(x)\\
=\;&\sum_{x,y,y^*}g^*(y^*,y,x)\pi(y^*|x)\frac{\partial}{\partial b}\pi_{b_0}(y|x)f_X(x)\\
=\;&\frac{\partial}{\partial b}p^*(\pi).
\end{aligned}
$$

This finishes the proof of equation (ii).

For equation (iii): Its LHS can be represented as

$$
\begin{aligned}
&\mathbb{E}\left\{\psi(w)\frac{\partial}{\partial\eta}\log l(w;\gamma_0,b_0,\eta_0)\right\}\\
=\;&\frac{1}{2}\sum_{x,y_1,y_2}\mathbb{E}_{z\sim\text{Ber}(g^*(y_1,y_2,x))}\Bigg\{\left(\frac{\pi(y_1|x)}{\pi_{\text{ref}}(y_1|x)}-\frac{\pi(y_2|x)}{\pi_{\text{ref}}(y_2|x)}\right)(z-g^*(y_1,y_2,x))\\
&\qquad\times\pi_{\text{ref}}(y_1|x)\pi_{\text{ref}}(y_2|x)\left(\frac{z}{g^*(y_1,y_2,x)}-\frac{1-z}{1-g^*(y_1,y_2,x)}\right)\frac{\partial}{\partial\eta}g_\eta(y_1,y_2,x)f_X(x)\Bigg\}\\
&+\frac{1}{2}\sum_{x,y_1,y_2,y^*}\mathbb{E}_{z\sim\text{Ber}(g^*(y_1,y_2,x))}\Bigg\{\left(g^*(y^*,y_1,x)+g^*(y^*,y_2,x)\right)\pi(y^*|x)\pi_{\text{ref}}(y_1|x)\\
&\qquad\times\pi_{\text{ref}}(y_2|x)f_X(x)\left(\frac{z}{g^*(y_1,y_2,x)}-\frac{1-z}{1-g^*(y_1,y_2,x)}\right)\frac{\partial}{\partial\eta}g_\eta(y_1,y_2,x).
\end{aligned}
$$

The second term is equal to zero due to the fact that

$$
\mathbb{E}_{z\sim\text{Ber}(g^*(y_1,y_2,x))}\left\{\frac{z}{g^*(y_1,y_2,x)}-\frac{1-z}{1-g^*(y_1,y_2,x)}\right\}=0.
$$

On the other hand, since

$$\begin{aligned}
&\mathbb{E}_{z\sim \mathrm{Ber}(g^*(y_1,y_2,x))}\left\{(z-g^*(y_1,y_2,x))\left(\frac{z}{g^*(y_1,y_2,x)}-\frac{1-z}{1-g^*(y_1,y_2,x)}\right)\right\}\\
=\;& g^*(y_1,y_2,x)\times(1-g^*(y_1,y_2,x))\frac{1}{g^*(y_1,y_2,x)}\\
&\quad+(1-g^*(y_1,y_2,x))\times(-g^*(y_1,y_2,x))\frac{-1}{1-g^*(y_1,y_2,x)}\\
=\;& 1,
\end{aligned}$$

the LHS in equation (iii) can be further represented by

$$\begin{aligned}
&\frac{1}{2}\sum_{x,y_1,y_2}\left(\frac{\pi(y_1|x)}{\pi_{\mathrm{ref}}(y_1|x)}-\frac{\pi(y_2|x)}{\pi_{\mathrm{ref}}(y_2|x)}\right)\pi_{\mathrm{ref}}(y_1|x)\pi_{\mathrm{ref}}(y_2|x)\frac{\partial}{\partial\eta}g_\eta(y_1,y_2,x)f_X(x)\\
=\;&\frac{1}{2}\sum_{x,y_1,y_2}\left(\pi(y_1|x)\pi_{\mathrm{ref}}(y_2|x)-\pi(y_2|x)\pi_{\mathrm{ref}}(y_1|x)\right)\frac{\partial}{\partial\eta}g_\eta(y_1,y_2,x)f_X(x)\\
=\;&\sum_{x,y_1,y_2}\pi(y_1|x)\pi_{\mathrm{ref}}(y_2|x)\frac{\partial}{\partial\eta}g_\eta(y_1,y_2,x)f_X(x)\\
=\;&\frac{\partial}{\partial\eta}p^*(\pi)
\end{aligned}\tag{18}$$

where the second-to-last equality follows from the fact $\frac{\partial}{\partial\eta}g_\eta(y_1,y_2,x)=-\frac{\partial}{\partial\eta}g_\eta(y_2,y_1,x)$. This finishes the proof of equation (iii).

Thus, with equation (i) - (iii) verified, Step 1 is proven.

Step 2: $\psi(W)-p^*(\pi)$ lies in the tangent space $\mathcal{T}_{\mathcal{M}}$.
By definition, the tangent space $\mathcal{T}_{\mathcal{M}}$ is the linear closure of the set of score functions of the all one-dimensional submodels regarding $\mathcal{M}$ that pass through true parameter; see Definition 2 in (Kallus & Uehara, 2020a). Based on the likelihood function in equation (17), we can explicitly calculate the tangent space of the data generating process model $\mathcal{M}$. In fact, the tangent space $\mathcal{T}_{\mathcal{M}}$ is a product space, which can be represented as $\mathcal{T}_f\bigoplus\mathcal{T}_\pi\bigoplus\mathcal{T}_g$, with $\mathcal{T}_f,\mathcal{T}_\pi,\mathcal{T}_g$ being the sets of score functions of all one-dimensional submodels passing through the marginal distribution $f_X(x)$, conditional distribution π_{ref} and preference probability g^*. Take the calculation of $\mathcal{T}_f$ as an example. Consider a one-dimensional submodel $\{f_\varepsilon(x)\}$, defined as

$$f_\varepsilon(x)=f(x)(1+\varepsilon q(x)),$$

where $q(x)$ satisfies $\sum_x f(x)q^2(x)<\infty$. Since we require f_ε to be a valid probability mass function, it must satisfy $\sum_x f_\varepsilon(x)=1$, which indicates $\mathbb{E}q(X)=0$. Then the score function with respect to ε is given by

$$\frac{d}{d\varepsilon}\log f_\varepsilon(x)=q(x).$$

Therefore, the tangent space for the marginal distribution function $f(x)$ can be represented as

$$\mathcal{T}_f=\left\{q(x):\mathbb{E}[q(X)]=0,\sum_x f(x)q^2(x)<\infty\right\}.$$

Following similar arguments, we can obtain

$$\begin{aligned}
\mathcal{T}_\pi &= \left\{q(y_1,x)+q(y_2,x):\mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[q(y,x)|X=x]=0,\sum_y\pi_{\mathrm{ref}}(y|x)q^2(y,x)<\infty\right\},\\
\mathcal{T}_g &= \left\{\frac{z-g^*(y_1,y_2,x)}{g^*(1-g^*)}q(y_1,y_2,x):\sum_{x,y_1,y_2}q^2(x,y_1,y_2)f(x)\pi_{\mathrm{ref}}(y_1|x)\pi_{\mathrm{ref}}(y_2|x)<\infty\right\}.
\end{aligned}$$

To verify $\psi(W) - p^*(\pi)$ lies in the tangent space, consider the following three functions:

$$\begin{aligned}
\psi_1(w) &:= \left(\frac{\pi(y_1|x)}{\pi_{\text{ref}}(y_1|x)} - \frac{\pi(y_2|x)}{\pi_{\text{ref}}(y_2|x)}\right)(z - g^*(y_1, y_2, x))\pi_{\text{ref}}(y_1|x)\pi_{\text{ref}}(y_2|x) f_X(x) \\
&= \frac{z - g^*(y_1, y_2, x)}{g^*(1-g^*)} g^*(1-g^*)\left(\pi(y_1|x)\pi_{\text{ref}}(y_2|x) - \pi(y_2|x)\pi_{\text{ref}}(y_1|x)\right) f_X(x), \\
\psi_2(y_1, y_2, x) &:= \mathbb{E}_{y^* \sim \pi}\left\{g(y^*, y_1, x) + g(y^*, y_2, x)\right\} - 2\mathbb{E}_{\substack{y \sim \pi_{\text{ref}(\bullet|x)} \\ y^* \sim \pi(\bullet|x)}}\left\{g(y^*, y, x)\right\}, \\
\psi_3(x) &:= 2\mathbb{E}_{\substack{y \sim \pi_{\text{ref}(\bullet|x)} \\ y^* \sim \pi(\bullet|x)}}\left\{g(y^*, y, x)\right\} - 2p^*(\pi).
\end{aligned}$$

It is easy to verify that $\psi_1(W) \in \mathcal{T}_g$, $\psi_2(Y^{(1)}, Y^{(2)}, X) \in \mathcal{T}_\pi$ and $\psi_3(X) \in \mathcal{T}_f$. Therefore,

$$\psi(W) - p^*(\pi) = \frac{1}{2}\Big(\psi_1(W) + \psi_2(Y^{(1)}, Y^{(2)}, X) + \psi_3(X)\Big) \in \mathcal{T}_{\mathcal{M}}.$$

This finishes the proof of Step 2.

With Step 1 and Step 2 verified, together with the fact that $\mathbb{E}\psi(W) = p^*(\pi)$, we obtain that $\psi(W)$ is an efficient influence function. $\square$

A.3. Proof of Theorem 2

Let $\mathbb{E}_n$ denote the empirical average over the n tuples $(X, Y^{(1)}, Y^{(2)}, Z)$ in the dataset $\mathcal{D}$. We further define the following norms:

$$\begin{aligned}
\|\widehat{g} - g^*\| &= \left(\mathbb{E}\left[\widehat{g}(Y^{(1)}, Y^{(2)}, X) - g^*(Y^{(1)}, Y^{(2)}, X)\right]^2\right)^{1/2} \\
\|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\| &= \max\left[\left(\mathbb{E}\max_y\left[\frac{\widehat{\pi}_{\text{ref}}(y|X)}{\pi_{\text{ref}}(y|X)} - 1\right]^2\right)^{1/2}, \left(\mathbb{E}\max_y\left[\frac{\pi_{\text{ref}}(y|X)}{\widehat{\pi}_{\text{ref}}(y|X)} - 1\right]^2\right)^{1/2}\right]
\end{aligned}$$

Accordingly, our estimator for $p^*(\pi)$ can be represented by $\mathbb{E}_n\psi(w; \pi, \widehat{\pi}_{\text{ref}}, \widehat{g})$. With some calculations, it can be further decomposed into

$$\mathbb{E}_n\psi(w; \pi, \widehat{\pi}_{\text{ref}}, \widehat{g}) = \mathbb{E}_n\psi(w; \pi, \pi_{\text{ref}}, g^*) + \text{I} + \text{II} + \text{III},$$

where

$$\begin{aligned}
\text{I} &= \frac{1}{2}\mathbb{E}_n\left\{\sum_{a=1}^{2}(-1)^a(Z - g^*(X, Y^{(1)}, Y^{(2)})\left[\frac{\pi(Y^{(a)}|X)}{\widehat{\pi}_{\text{ref}}(Y^{(a)}|X)} - \frac{\pi(Y^{(a)}|X)}{\pi_{\text{ref}}(Y^{(a)}|X)}\right]\right\}, \\
\text{II} &= \frac{1}{2}\mathbb{E}_n\left\{\sum_{a=1}^{2}\mathbb{E}_{y\sim\pi(\bullet|x)}\left[(\widehat{g} - g^*)(X, y, Y^{(a)})\right]\right\} \\
&\quad -\frac{1}{2}\mathbb{E}_n\left\{\sum_{a=1}^{2}(-1)^a\frac{\pi(Y^{(a)}|X)}{\pi_{\text{ref}}(Y^{(a)}|X)}(\widehat{g} - g^*)(X, Y^{(1)}, Y^{(2)})\right\}, \\
\text{III} &= \frac{1}{2}\mathbb{E}_n\left\{(-1)^a(\widehat{g} - g^*)(X, Y^{(1)}, Y^{(2)})\left[\frac{\pi(Y^{(a)}|X)}{\widehat{\pi}_{\text{ref}}(Y^{(a)}|X)} - \frac{\pi(Y^{(a)}|X)}{\pi_{\text{ref}}(Y^{(a)}|X)}\right]\right\}.
\end{aligned}$$

From Lemma A8, we know that $\mathbb{E}_n\psi(w; \pi, \pi_{\text{ref}}, g^*)$ is an unbiased estimator for $p^*(\pi)$ with variance equal to SEB. Since both $\widehat{\pi}_{\text{ref}}$ and $\widehat{g}$ are obtained from external models independent of $\mathcal{D}$, analogous to the proof of Lemma 1, we know that the first term I and the second term II have zero means. The third term III is the bias term. Therefore, we obtain the following bias-variance decomposition for MSE($\widehat{p}_{\text{DR}}$):

$$\text{MSE}(\widehat{p}_{\text{DR}}(\pi)) = \text{Var}(\mathbb{E}_n\psi(w; \pi, \pi_{\text{ref}}, g^*) + \text{I} + \text{II} + \text{III}) + (\mathbb{E}[\text{III}])^2 \tag{19}$$

Since g^* is bounded by 1, under the coverage assumption, we obtain that

$$\begin{aligned}\text{Var}(\mathbb{E}_n\psi(w;\pi,\pi_{\text{ref}},g^*) &= \frac{1}{n}\text{Var}(\psi(w;\pi,\pi_{\text{ref}},g^*)) = O\Big(\frac{1}{n}\mathbb{E}\frac{\pi^2(Y|X)}{\pi^2_{\text{ref}}(Y|X)}\Big)\\ &= O\Big(\frac{1}{n}\sum_y\frac{\pi^2(y|X)}{\pi_{\text{ref}}(y|X)}\Big) = O\left(\frac{1}{n\epsilon}\right).\end{aligned}\tag{20}$$

Moreover, we have

$$\begin{aligned}\mathbb{E}\text{I}^2 =&\frac{1}{2n}\mathbb{E}\left\{(Z-g^*(X,Y^{(1)},Y^{(2)})^2\left[\frac{\pi(Y|X)}{\widehat{\pi}_{\text{ref}}(Y|X)}-\frac{\pi(Y|X)}{\pi_{\text{ref}}(Y|X)}\right]^2\right\}\\ \leq&\frac{1}{2n}\mathbb{E}\left\{\frac{\pi^2(Y|X)}{\pi^2_{\text{ref}}(Y|X)}\left[\frac{\pi_{\text{ref}}(Y|X)}{\widehat{\pi}_{\text{ref}}(Y|X)}-1\right]^2\right\}\\ =&\frac{1}{2n}\mathbb{E}_X\left\{\sum_y\frac{\pi^2(y|X)}{\pi_{\text{ref}}(y|X)}\left[\frac{\pi_{\text{ref}}(y|X)}{\widehat{\pi}_{\text{ref}}(y|X)}-1\right]^2\right\}\\ =&O\left(\frac{1}{n\epsilon}\left\|\frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}}-1\right\|^2\right).\end{aligned}\tag{21}$$

Follow a similar argument, we obtain

$$\mathbb{E}\text{II}^2 \quad=\quad O\left(\frac{1}{n\epsilon}\|\widehat{g}-g^*\|^2\right),\qquad \mathbb{E}\text{III}^2 = O\left(\frac{1}{n\epsilon}\left\|\frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}}-1\right\|^2\right).\tag{22}$$

By Cauchy inequality, we have for any random variables X and Y that $|\text{Cov}(X,Y)|\leq\sqrt{\text{Var}(X)\text{Var}(Y)}$. It follows that

$$\begin{aligned}\text{Cov}\left(\mathbb{E}_n\psi(w;\pi,\pi_{\text{ref}},g^*),\text{I}+\text{III}\right) =&O\left(\frac{1}{n\epsilon}\left\|\frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}}-1\right\|\right),\\ \text{Cov}\left(\mathbb{E}_n\psi(w;\pi,\pi_{\text{ref}},g^*),\text{II}\right) =&O\left(\frac{1}{n\epsilon}\|\widehat{g}-g^*\|\right),\\ \text{Cov}\left(\text{I}+\text{III},\text{II}\right) =&O\left(\frac{1}{n\epsilon}\|\widehat{g}-g^*\|\cdot\left\|\frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}}-1\right\|\right).\end{aligned}\tag{23}$$

Assuming that $\|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}}-1\|$ is bounded, the high-order terms Var(I), Var(I) and Var(III) are dominated by the first two terms in (23). Combining equations (20), (21),(22) and (23) yields

$$\text{Var}(\mathbb{E}_n\psi(w;\pi,\pi_{\text{ref}},g^*)+\text{I}+\text{II}+\text{III}) = \text{SEB}+O\left(\frac{1}{n\epsilon}\|\widehat{g}-g^*\|\right)+O\left(\frac{1}{n\epsilon}\left\|\frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}}-1\right\|\right).\tag{24}$$

Finally, using Cauchy inequality again, we obtain that

$$\begin{aligned}\mathbb{E}\big|\text{III}\big| \quad&=\quad O\left(\mathbb{E}\left\{(\widehat{g}-g^*)^2(X,Y^{(1)},Y^{(2)})\right\}^{1/2}\mathbb{E}\left\{\left[\frac{\pi_{\text{ref}}(Y|X)}{\widehat{\pi}^2_{\text{ref}}(Y|X)}-1\right]^2\frac{\pi^2(Y|X)}{\pi^2_{\text{ref}}(Y|X)}\right\}^{1/2}\right)\\ &=\quad O\left(\frac{1}{\sqrt{\epsilon}}\|\widehat{g}-g^*\|\cdot\|\pi_{\text{ref}}/\widehat{\pi}_{\text{ref}}-1\|\right).\end{aligned}$$

Combining (19) and (25), we obtain

$$\begin{aligned}\text{MSE}(\widehat{p}_{\text{DR}}(\pi)) \quad&=\quad \mathbb{E}\left\{\mathbb{E}_n\psi(w;\pi,\widehat{\pi}_{\text{ref}},\widehat{g})-p^*(\pi)\right\}^2\\ &=\quad \text{SEB}+O\left(\frac{1}{n\epsilon}\|\widehat{g}-g^*\|\right)+O\left(\frac{1}{n\epsilon}\|\pi_{\text{ref}}/\widehat{\pi}_{\text{ref}}-1\|\right)\\ &\qquad +O\left(\frac{1}{\epsilon}\|\pi_{\text{ref}}/\widehat{\pi}_{\text{ref}}-1\|^2\cdot\|\widehat{g}-g^*\|^2\right).\end{aligned}$$

This finishes the proof of Theorem 2.

A.4. Proofs of Corollaries 3 and 4

The proofs of Corollaries 3 and 4 follow directly from the assertion of Theorem 2.

A.5. Proof of Theorem 5

Let π^* denote the maximizer of $p^*(\pi)$ in the policy class Π. Throughout the proof, for any policies π_1 and π_2, we use a shorthand and write $\mathbb{E}_{X\sim\mathcal{D}} D_{\mathrm{KL}}[\pi_1(\bullet \mid X) \,\|\, \pi_2(\bullet \mid X)]$ as $\mathrm{KL}(\pi_1\|\pi_2)$. Since $\widehat{\pi}$ is a maximizer of $\widehat{p}_{\mathrm{DR}}(\pi) - \beta\mathrm{KL}(\pi\|\widehat{\pi}_{\mathrm{ref}})$, we have

$$\widehat{p}_{\mathrm{DR}}(\widehat{\pi}) - \beta\mathrm{KL}(\widehat{\pi}\|\widehat{\pi}_{\mathrm{ref}}) \geq \widehat{p}_{\mathrm{DR}}(\pi^*) - \beta\mathrm{KL}(\pi^*\|\widehat{\pi}_{\mathrm{ref}}).$$

It directly follows that

$$\begin{aligned}
& p^*(\pi^*) - p^*(\widehat{\pi}) \\
\leq\ & p^*(\pi^*) - \widehat{p}_{\mathrm{DR}}(\pi^*) + \widehat{p}_{\mathrm{DR}}(\widehat{\pi}) - p^*(\widehat{\pi}) + \beta(\mathrm{KL}(\pi^*\|\widehat{\pi}_{\mathrm{ref}}) - \mathrm{KL}(\widehat{\pi}\|\widehat{\pi}_{\mathrm{ref}})) \\
\leq\ & \mathbb{E}\,|p^*(\pi^*) - \widehat{p}_{\mathrm{DR}}(\pi^*)| + \mathbb{E}\,|\widehat{p}_{\mathrm{DR}}(\widehat{\pi}) - p^*(\widehat{\pi})| + O\left(\beta\log^{-1}\epsilon\right) \\
\leq\ & 2\mathbb{E}\sup_{\pi\in\Pi}|p^*(\pi) - \widehat{p}_{\mathrm{DR}}(\pi)| + O\left(\beta\log^{-1}\epsilon\right),
\end{aligned} \tag{25}$$

where the second inequality follows from the coverage assumption that $\mathrm{KL}(\pi\|\widehat{\pi}_{\mathrm{ref}}) = \mathbb{E}_{X\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(\bullet|X)}\log\frac{\pi(y|X)}{\widehat{\pi}_{\mathrm{ref}}(y|X)} = O(\log^{-1}\epsilon)$.

Additionally, following the proof of Theorem 2, the bias of the proposed preference evaluation estimator can be upper bounded by

$$\sup_{\pi\in\Pi}|\mathbb{E}[p^*(\pi) - \widehat{p}_{\mathrm{DR}}(\pi)]| = \mathbb{E}|\mathbb{E}_n\psi(w;\pi,\pi_{\mathrm{ref}},g^*) - p^*(\pi)| + O\left(\frac{1}{\epsilon}\|\widehat{g} - g^*\|\cdot\|\widehat{\pi}_{\mathrm{ref}}/\pi_{\mathrm{ref}} - 1\|\right). \tag{26}$$

It remains to upper bound the empirical process term $\mathbb{E}\sup_{\pi\in\Pi}|\widehat{p}_{\mathrm{DR}}(\pi) - \mathbb{E}\widehat{p}_{\mathrm{DR}}(\pi)|$. Toward that end, we employ Corollary 5.1 in (Chernozhukov et al., 2014). To invoke this corollary, notice that

1. According to Assumption 4, Π is a policy class with VC dimension v. Under Assumption 1, it follows from Lemma A.6 in (Chernozhukov et al., 2014) that the function class $\mathcal{F} = \{\psi(\bullet,\pi,\widehat{\pi}_{\mathrm{ref}},\widehat{g})|\pi\in\Pi\}$ also has a VC dimension of v.
2. Using the coverage assumption again, the function class $\mathcal{F}$ is uniformly bounded by $O(1/\epsilon)$.
3. The variance $\sup_{f\in\mathcal{F}}\mathrm{Var}(f(W))$ is uniformly bounded by $O(1/\epsilon^2)$.

Consequently, an application of Corollary 5.1 in (Chernozhukov et al., 2014) yields that

$$\begin{aligned}
\mathbb{E}\sup_{\pi\in\Pi}|\widehat{p}_{\mathrm{DR}}(\pi) - \mathbb{E}[\widehat{p}_{\mathrm{DR}}(\pi)]| &= O\left(\frac{1}{\sqrt{n}}\sqrt{\frac{v}{\epsilon^2}\log^{-1}\epsilon^2} + \frac{v}{n}\log^{-1}\epsilon^2\right) \\
&= O\left(\frac{1}{\epsilon}\sqrt{\frac{v\log^{-1}\epsilon}{n}} + \frac{v\log^{-1}\epsilon}{n\epsilon}\right).
\end{aligned}$$

Combining equations (25), (26) and (27), we obtain for any $\pi\in\Pi$ that

$$p^*(\pi^*) - p^*(\widehat{\pi}) = O\left(\beta\log^{-1}\epsilon + \frac{1}{\epsilon}\sqrt{\frac{v\log^{-1}\epsilon}{n}} + \frac{v\log^{-1}\epsilon}{n\epsilon} + \frac{1}{\epsilon}\|\widehat{g} - g^*\|\cdot\|\widehat{\pi}_{\mathrm{ref}}/\pi_{\mathrm{ref}} - 1\|\right).$$

This completes the proof of Theorem 5.

A.6. Proof of Corollary 6

The proof of Corollary 6 follows directly from the assertion of Theorem 5.

A.7. Proof of Theorem 7

Suboptimality gap for DRPO: If the BT assumption holds, we have $g^*(y_1, y_2, x) = \sigma(r^*(y_1, x) - r^*(y_2, x))$ where $\sigma(x) = 1/(1+e^{-x})$ is the sigmoid function. Since the sigmoid function is monotonically increasing, under the realizability assumption, π^* which maximizes $J(\pi)$ also maximizes $p^*(\pi)$. This follows from the classical results on the maximum rank correlation estimator that has been widely studied in the econometrics literature (see e.g., **??**). Therefore,

$$\begin{aligned} p^*(\pi^*) - p^*(\widehat{\pi}) &= \mathbb{E}_{y^*\sim\pi^*, \widetilde{y}\sim\widehat{\pi}, y\sim\pi_{\text{ref}}} \{g^*(y^*, y, x) - g^*(\widetilde{y}, y, x)\} \\ &= \mathbb{E}_{y^*\sim\pi^*, \widetilde{y}\sim\widehat{\pi}, y\sim\pi_{\text{ref}}} \{\sigma'(\xi)\,[(r^*(y^*, x) - r^*(y, x)) - (r^*(\widetilde{y}, x) - r^*(y, x))]\}^2 \\ &= \mathbb{E}_{y^*\sim\pi^*, \widetilde{y}\sim\widehat{\pi}} \{\sigma'(\xi)(r^*(y^*, x) - r^*(\widetilde{y}, x))\} \\ &\geq C_0(J(\pi^*) - J(\widehat{\pi})), \end{aligned}$$

where C_0 is some positive constant and ξ is some real number between $r^*(y^*, x) - r^*(y, x)$ and $r^*(\widetilde{y}, x) - r^*(y, x)$. Here, the second equality follows from mean value theorem. The last equality follows from the identity that $\sigma'(x) = \sigma(x)(1 - \sigma(x))$, which is bounded away from zeroo under Assumption 2 that the reward is bounded by some constant. Thus, we obtain $J(\pi^*) - J(\widehat{\pi}) = O(\text{Reg}(\widehat{\pi}))$ and the suboptimality gap for DRPO follows directly from the assertion in Theorem 5.

Suboptimality gap for PPO-based algorithm: We begin with some notations. For a given estimated reward $\widehat{r}$, define

- $l(\pi) = \mathbb{E}[\mathbb{E}_{y\sim\pi}\widehat{r}(y, X)] - \beta\text{KL}(\pi\|\pi_{\text{ref}})$,
- $l_n(\pi) = \mathbb{E}_n\mathbb{E}_{y\sim\pi}\widehat{r}(y, X) - \beta\text{KL}(\pi\|\pi_{\text{ref}})$,
- $\widetilde{\pi} = \arg\max_{\pi\in\Pi} l(\pi)$,
- $\widehat{\pi} = \arg\max_{\pi\in\Pi} l_n(\pi)$.

Using the fact that $l(\widetilde{\pi}) \geq l(\pi^*)$ and $l_n(\widehat{\pi}) \geq l_n(\widetilde{\pi})$, we obtain the following upper bound:

$$\begin{aligned} J(\pi^*) - J(\widehat{\pi}) &\leq \mathbb{E}\{[J(\pi^*) - l(\pi^*)] + [l(\widetilde{\pi}) - l_n(\widetilde{\pi})] + [l_n(\widehat{\pi}) - l(\widehat{\pi})] + [l(\widehat{\pi}) - J(\widehat{\pi})]\} \\ &\leq \mathbb{E}\{[J(\pi^*) - l(\pi^*)]\} + \mathbb{E}\{[l(\widehat{\pi}) - J(\widehat{\pi})]\} + 2\mathbb{E}\sup_{\pi\in\Pi}\{|l(\pi) - l_n(\pi)|\}. \end{aligned} \tag{27}$$

For the first term, we have

$$\begin{aligned} \mathbb{E}\{|J(\pi^*) - l(\pi^*)|\} &= \mathbb{E}_{y\sim\pi^*}|\widehat{r}(y, X) - r^*(y, X)| + \beta\text{KL}(\pi^*\|\pi_{\text{ref}}) \\ &= \mathbb{E}_{y\sim\pi_{\text{ref}}}\left[\frac{\pi^*(y|X)}{\pi_{\text{ref}}(y|X)}|\widehat{r}(y, X) - r^*(y, X)|\right] + O(\beta\log^{-1}\epsilon) \\ &= O\left(\frac{1}{\sqrt{\epsilon}}\|\widehat{r} - r^*\|\right) + O(\beta\log^{-1}\epsilon), \end{aligned} \tag{28}$$

where the last equation follows from Cauchy inequality.

Using a similar argument, we obtain that $\mathbb{E}\{|l(\widehat{\pi}) - J(\widehat{\pi})|\} = O\left(\frac{1}{\sqrt{\epsilon}}\|\widehat{r} - r^*\| + \beta\log^{-1}\epsilon\right)$.

Finally, under assumption 2, the function class $\mathcal{F} = \left\{\sum_y \widehat{r}(y, X)\pi(y|X)\middle|\pi\in\Pi\right\}$ is bounded by a constant. Using similar arguments to the proof of Theorem 5, we can employ Corollary 5.1 in (Chernozhukov et al., 2014) to show that

$$\mathbb{E}\sup_{\pi\in\Pi}\{|l(\pi) - l_n(\pi)|\} = O\left(\frac{v}{n} + \sqrt{\frac{v}{n}}\right) + O(\beta\log^{-1}\epsilon). \tag{29}$$

Combining equations (27), (28) and (29), we obtain that

$$J(\pi^*) - J(\widehat{\pi}) = O\left(\beta\log^{-1}\epsilon + \frac{v}{n} + \sqrt{\frac{v}{n}} + \frac{1}{\sqrt{\epsilon}}\|\widehat{r} - r^*\|\right).$$

Suboptimality gap for DPO-based algorithm: We need some additional technical conditions to prove the suboptimality gap for DPO-based algorithms. Recall that when BT-model holds, there exists a one-on-one correspondence between the policy and reward model (Rafailov et al., 2023). We further assume

Assumption 5 (Realizability)**.** The oracle reward r^* lies in the bounded reward function class $\mathcal{R} = \{\beta \log(\pi(y|x)/\pi_{\text{ref}}(y|x)) + \beta Z(x) : \pi \in \Pi\}$ induced by the policy class Π.

Assumption 6 (Coverage)**.** Both π_{ref} and $\widehat{\pi}_{\text{ref}}$ are lower bounded by some constant $\epsilon > 0$.

Assumption 7 (Suboptimality gap for oracle reward)**.** Let $y_x^* = \arg\max_y r^*(y|x)$ and $\bar{y}_x = \arg\max_{y \neq y^*} r^*(y|x)$. There exists a positive constant $\bar{c}$ such that for any x,

$$r^*(y_x^*, x) - r^*(\bar{y}_x, x) \geq \bar{c}.$$

Notice that both the realizability and the coverage in Assumptions 5 and 6 differ from those in the main text. Specifically, Assumption 5 imposes the realizability assumption on the oracle reward rather than the optimal policy whereas Assumption 6 is stronger than that in the main text by requiring the denominators of the IS ratios to be strictly positive.

We next introduce some notations. For a given estimated reference policy $\widehat{\pi}_{\text{ref}}$, any policy π induce a reward function

$$r^{\pi}(y, x) = \beta \log\left(\frac{\pi(y|x)}{\widehat{\pi}_{\text{ref}}(y|x)}\right) + \beta Z(x) \tag{30}$$

Let $l(\pi)$ be the log-likelihood function induced by reward r^{π} and $l^*(\pi)$ be its variant with $\widehat{\pi}_{\text{ref}}$ in the denominator of (30) replaced by the ground truth π_{ref}. Denote $\widetilde{\pi} = \arg\max_{\pi} \mathbb{E}_n l(\pi)$ and $\widehat{\pi} = \arg\max_{\pi} \mathbb{E} l(\pi)$. It follows that

$$\begin{aligned}
& \mathbb{E}_n l(\widetilde{\pi}) - \mathbb{E}_n l(\widehat{\pi}) - \mathbb{E} l(\widetilde{\pi}) + \mathbb{E} l(\widehat{\pi}) \\
\leq\; & \mathbb{E} l(\widehat{\pi}) - \mathbb{E} l(\widetilde{\pi}) \\
\leq\; & \mathbb{E} l(\widehat{\pi}) - \mathbb{E} l^*(\widetilde{\pi}) + \mathbb{E} l^*(\widetilde{\pi}) - \mathbb{E} l(\widetilde{\pi}) \\
\leq\; & -C_1 \mathbb{E} \left\| \widehat{r}(y_1, x) - \widehat{r}(y_2, x) - r^*(y_1, x) + r^*(y_2, x) \right\|_2^2 + \beta^2 C_2 \mathbb{E} \left(\log \frac{\widehat{\pi}_{\text{ref}}(Y^{(1)}|X)}{\pi_{\text{ref}}(Y^{(1)}|X)} \right)^2 \\
\leq\; & -C_1 \sigma^2 + \beta^2 C_2 \|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\|^2,
\end{aligned} \tag{31}$$

where $\sigma^2 = \mathbb{E} \left\| \widehat{r}(y_1, x) - \widehat{r}(y_2, x) - r^*(y_1, x) + r^*(y_2, x) \right\|_2^2$, both C_1 and C_2 are positive constants because the Hessian matrix is bounded away from zero and infinity, which follows from the boundedness assumption on the reward, and the last inequality is due to that $\mathbb{E} \left\{ \log \frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}} \right\}^2 \leq \mathbb{E} \left(\frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}} - 1 \right)^2$, which can be further bounded by $\| \frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}} - 1 \|_2^2$, according to the definition of the norm.

Moreover, according to Corollary 5.1 in (Chernozhukov et al., 2014), using similar arguments to the proof of Theorem 5 and PPO-based algorithms, we have

$$\begin{aligned}
\mathbb{E}_n l(\widetilde{\pi}) - \mathbb{E}_n l(\widehat{\pi}) - \mathbb{E} l(\widetilde{\pi}) + \mathbb{E} l(\widehat{\pi}) \quad & \leq \quad 2\mathbb{E} \sup_{\pi \in \Pi} |l(\pi) - \mathbb{E} l_n(\pi)| \\
& \leq \quad O(\sigma \sqrt{\frac{v}{n}} + \frac{v}{n}).
\end{aligned} \tag{32}$$

This together with equation (31) yields that $C_1(\sigma - \bar{c}\sqrt{v/n})^2 \leq \bar{c}v/n + \beta^2 C_2 \|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\|^2$ for some constant $\bar{c} > 0$, and hence

$$\sigma = O\left(\sqrt{\frac{v}{n}} + \beta \|\pi_{\text{ref}} - \widehat{\pi}_{\text{ref}}\|_2 \right). \tag{33}$$

Recall that π^* is the true optimal policy, and $\widehat{\pi}$ in this part of the proof denotes DPO's estimated optimal policy. We further define $\widehat{\pi}^*$ as a softmax optimal policy based on the oracle reward function r^*

$$\widehat{\pi}^*(y|x) = \frac{\widehat{\pi}^*(y|x) \exp(\frac{1}{\beta} r^*(y, x))}{\sum_{y'} \widehat{\pi}^*(y'|x) \exp(\frac{1}{\beta} r^*(y', x))}.$$

With some calculations, it follows that

$$\begin{aligned}
& J(\pi^*) - J(\widehat{\pi}) \\
=& \mathbb{E}[\mathbb{E}_{y \sim \pi^*} r^*(y, X) - \mathbb{E}_{y \sim \widehat{\pi}} r^*(y, X)] \\
=& \mathbb{E}(\mathbb{E}_{y \sim \pi^*} r^*(y, X) - \mathbb{E}_{y \sim \widehat{\pi}^*} r^*(y, X)) + \mathbb{E}(\mathbb{E}_{y \sim \widehat{\pi}^*} r^*(y, X) - \mathbb{E}_{y \sim \widehat{\pi}} r^*(y, X)),
\end{aligned} \tag{34}$$

where the outer expectations are taken with respect to the prompt distribution.

Recall that y_x^* denotes the optimal response to the prompt x. The first term $\mathbb{E}[\mathbb{E}_{y\sim\pi^*} r^*(y,X) - \mathbb{E}_{y\sim\widehat{\pi}^*} r^*(y,X)]$ can be upper bounded by

$$\begin{aligned}
\mathbb{E}r^*(y_X^*,X) - \mathbb{E}[\mathbb{E}_{y\sim\widehat{\pi}^*} r^*(y,X)] &= \mathbb{E}r^*(y_X^*,X) - \mathbb{E}\left\{\frac{\sum_y r^*(y,X)\widehat{\pi}_{\text{ref}}(y|X)\exp\left(\frac{1}{\beta}r^*(y,X)\right)}{\sum_y \widehat{\pi}_{\text{ref}}(y|X)\exp\left(\frac{1}{\beta}r^*(y,X)\right)}\right\} \\
&\le \mathbb{E}r^*(y_X^*,X) - \mathbb{E}\left\{\frac{r^*(y_X^*,X)\widehat{\pi}_{\text{ref}}(y_X^*|X)\exp\left(\frac{1}{\beta}r^*(y_X^*,X)\right)}{\sum_y \widehat{\pi}_{\text{ref}}(y|X)\exp\left(\frac{1}{\beta}r^*(y,X)\right)}\right\} \\
&= O\left(\frac{1}{\epsilon}\exp\left(-\frac{\bar{c}}{\beta}\right)\right),
\end{aligned}$$

where the last equality is due to that under Assumptions 6 and 7, the difference between 1 and the ratio $\frac{\widehat{\pi}_{\text{ref}}(y_X^*|X)\exp\left(\frac{1}{\beta}r^*(y_X^*,X)\right)}{\sum_y \widehat{\pi}_{\text{ref}}(y|X)\exp\left(\frac{1}{\beta}r^*(y,X)\right)}$ is of the order $O\left(\frac{1}{\epsilon}\exp\left(-\frac{\bar{c}}{\beta}\right)\right)$, almost surely.

Using mean value theorem, the second term can be bounded by

$$\mathbb{E}\sum_y |\widehat{\pi}(y|X) - \widehat{\pi}^*(y|X)| \le \frac{1}{\beta}\mathbb{E}\max_y |\widehat{r}(y,X) - r^*(y,X)| \le \frac{1}{\beta\sqrt{\epsilon}}\|\widehat{r} - r^*\|_2, \tag{35}$$

where the last inequality follows from the fact that

$$\begin{aligned}
\|\widehat{r} - r^*\|_2 &= \mathbb{E}\{(\widehat{r} - r^*)^2\}^{1/2} \\
&= \mathbb{E}\left\{\sum_y \pi_{\text{ref}}(y|X)(\widehat{r}(y|X) - r^*(y|X))^2\right\}^{1/2} \\
&\ge \sqrt{\epsilon}\mathbb{E}\left\{\sum_y (\widehat{r}(y|X) - r^*(y|X))^2\right\}^{1/2} \\
&\ge \sqrt{\epsilon}\max_y |\widehat{r}(y,X) - r^*(y|X)|. \qquad (36)
\end{aligned}$$

To complete the proof, it remains to upper bound $\|\widehat{r} - r^*\|_2$ using σ^2. Recall that $\sigma^2 = \mathbb{E}\left\|\widehat{r}(Y^{(1)},X) - \widehat{r}(Y^{(2)},X) - r^*(Y^{(1)},X) + r^*(Y^{(2)},X)\right\|_2^2$. Since $Y^{(2)}$ is independent of $Y^{(1)}$ given X and that π_{ref} is lower bounded by $\epsilon > 0$, it follows that

$$\sigma^2 \ge \epsilon\mathbb{E}\left\|\widehat{r}(Y^{(1)},X) - \widehat{r}(y_0,X) - r^*(Y^{(1)},X) + r^*(y_0,X)\right\|_2^2,$$

for a fixed y_0. Notice that the RHS corresponds to the mean squared error between $\widehat{r}$ and r^*, up to a baseline term that is independent of $Y^{(1)}$. Without loss of generality, we can assume this baseline term $r^*(y_0,X) - \widehat{r}(y_0,X)$ this equal to zero without affecting the validity of the proof. This is because the true reward can be redefined as $r^*(\bullet,X) - r^*(y_0,X)$, since it is equivalent up to a function independent of the response. Similarly, the estimated optimal policy $\widehat{\pi}(\bullet|x)$ computed by DPO can be represented using the difference $\widehat{r}(\bullet,x) - \widehat{r}(y_0,x)$, and we can replace $\widehat{r}$ in (35) using this difference. Consequently, we obtain that $\sigma^2 \ge \epsilon\|\widehat{r} - r^*\|^2$ and hence

$$\|\widehat{r} - r^*\| = O\left(\epsilon^{-1/2}\sqrt{\frac{v}{n}} + \beta\epsilon^{-1/2}\|\pi_{\text{ref}} - \widehat{\pi}_{\text{ref}}\|_2\right).$$

Combining this together with equations (33) and (34), we obtain that the regret is upper bounded by

$$O\left(\frac{\exp(-\bar{c}\beta^{-1})}{\epsilon} + \frac{1}{\beta\epsilon}\sqrt{\frac{v}{n}} + \frac{1}{\epsilon}\|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\|\right).$$

The proof is hence completed.

B. DRPO Algorithm Details and Practical Implementation

This section details our proposed algorithm. Notably, the reference model $\widehat{\pi}_{\text{ref}}$ and the preference model $\widehat{g}$ are pre-trained independently prior to policy optimization. The proposed objective function is defined as

$$\mathcal{J}(\pi_\theta; \widehat{\pi}_{\text{ref}}, \widehat{g}_\eta, \mathcal{D}) = \widehat{p}_{\text{DR}}(\pi) - \beta \mathbb{E}_{X\sim\mathcal{D}} D_{\text{KL}}[\pi(\bullet \mid X) \,\|\, \widehat{\pi}_{\text{ref}}(\bullet \mid X)]. \tag{37}$$

The gradient of $\mathcal{J}(\pi_\theta)$ is given by:

$$\begin{aligned} \nabla_\theta \mathcal{J}(\pi_\theta) \quad = \quad & \frac{1}{2}\mathbb{E}_{X,Y^{(1)},Y^{(2)}\sim\mathcal{D}} \Bigg\{ \sum_{a=1}^{2} \mathbb{E}_{y\sim\pi_\theta(\bullet|X)} \left[\widehat{g}(X, y, Y^{(a)}) \nabla_\theta \log \pi_\theta(y|X) \right] \\ & + \sum_{a=1}^{2} (-1)^{a-1} \frac{\nabla_\theta \pi_\theta(Y^{(a)}|X)}{\widehat{\pi}_{\text{ref}}(Y^{(a)}|X)} \left(Z - \widehat{g}(X, Y^{(1)}, Y^{(2)}) \right) \Bigg\} \\ & - \beta \nabla_\theta D_{\text{KL}}[\pi_\theta(\bullet \mid X) \,\|\, \widehat{\pi}_{\text{ref}}(\bullet \mid X)] \end{aligned} \tag{38}$$

Intuitively, the gradient operates as follows: The first term guides the policy to favor responses preferred by the preference model $\widehat{g}$. When $Y^{(1)} \succ Y^{(2)}$, which means $Z = 1$, the second term enhances the likelihood of $Y^{(1)}$ while diminishing the likelihood of $Y^{(2)}$, and vice versa.

The empirical loss function is constructed such that its negative gradient corresponds to $\nabla_\theta \mathcal{J}(\pi_\theta)$ in Equation 38. The direct-method term is approximated using Monte Carlo sampling by drawing several new responses $\mathcal{D}_X^* := \{Y^* \mid Y^* \sim \pi_\theta(\bullet \mid X)\}$ from the current policy π_θ for a given prompt X at each policy update. A k3-type empirical KL divergence is utilized, following (Shao et al., 2024).

$$\begin{aligned} \mathcal{L}_{\text{DRPO}} \quad = \quad & -\frac{1}{2}\mathbb{E}_{X,Y^{(1)},Y^{(2)}\sim\mathcal{D}} \Bigg\{ \mathbb{E}_{Y^*\sim\mathcal{D}_X^*} \left[\sum_{a=1}^{2} \widehat{g}(Y^*, Y^{(a)}, X) \log \pi_\theta(Y^*|X) \right] \\ & + \sum_{a=1}^{2} (-1)^{a-1} \frac{\pi_\theta(Y^{(a)}|X)}{\pi_{\text{ref}}(Y^{(a)}|X)} \left(Z - \widehat{g}(X, Y^{(1)}, Y^{(2)}) \right) \Bigg\} \\ & + \beta \mathbb{E}_{Y^*\sim\mathcal{D}_X^*, X\sim\mathcal{D}} \left[\frac{\widehat{\pi}_{\text{ref}}(Y^* \mid X)}{\pi_\theta(Y^* \mid X)} - 1 - \log \frac{\widehat{\pi}_{\text{ref}}(Y^* \mid X)}{\pi_\theta(Y^* \mid X)} \right] \end{aligned} \tag{39}$$

Maximization of $\mathcal{J}(\pi_\theta)$ is achieved by minimizing the loss function. In practice, the original offline dataset is augmented to $\mathcal{D}$ to $\widetilde{\mathcal{D}}$ by including swapped pairs (i.e. for $(X, Y^{(1)}, Y^{(2)}, Z)$, we add $(X, Y^{(2)}, Y^{(1)}, 1 - Z)$ to $\widetilde{\mathcal{D}}$, simplifying the empirical loss function (39). Furthermore, the importance sampling ratio is clipped, and its calculation is decoupled from the gradient computation. This is achieved by stopping auto-differentiation for the ratio and multiplying the importance sampling term by $\log \pi_\theta$, which shrinks (rather than eliminates) gradients in small $\widehat{\pi}_{\text{ref}}$ regions while maintaining approximate arithmetic equivalence. Consequently, the loss function is reformulated as:

$$\begin{aligned} \mathcal{L}_{\text{DRPO}} \quad = \quad & -\frac{1}{2}\mathbb{E}_{X,Y^{(1)},Y^{(2)}\sim\widetilde{\mathcal{D}}} \Bigg\{ \underbrace{\mathbb{E}_{Y^*\sim\mathcal{D}_X^*} \left[\widehat{g}(Y^*, Y^{(2)}, X) \log \pi_\theta(Y^*|X) \right]}_{\text{term I}} \\ & + \text{sg}\Bigg(\underbrace{\text{clip}\Big(\frac{\pi_\theta(Y^{(1)}|X)}{\pi_{\text{ref}}(Y^{(1)}|X)}, 1 - \epsilon_1, 1 + \epsilon_2 \Big) \left(Z - \widehat{g}(Y^{(1)}, Y^{(2)}, X) \right)}_{\text{term II}} \Bigg) \log \pi_\theta(Y^{(1)} \mid X) \Bigg\} \\ & + \beta \mathbb{E}_{Y^*\sim\mathcal{D}_X^*, X\sim\widetilde{\mathcal{D}}} \left[\frac{\widehat{\pi}_{\text{ref}}(Y^* \mid X)}{\pi_\theta(Y^* \mid X)} - 1 - \log \frac{\widehat{\pi}_{\text{ref}}(Y^* \mid X)}{\pi_\theta(Y^* \mid X)} \right] \end{aligned} \tag{40}$$

where $\text{sg}(\bullet)$ denotes stop-gradient operation, $\text{clip}(\bullet, a, b)$ signifies clipping values to the interval $[a, b]$, and ϵ_1, ϵ_2 are hyperparameters defining the clipping range. See full details in Algorithm 1

Algorithm 1 Double Robust Preference Optimization

```
Require: reference policy $\widehat{\pi}_{\mathrm{ref}}$, preference model $\widehat{g}$, offline dataset $\widetilde{\mathcal{D}} = \{X_i, Y_i^{(1)}, Y_i^{(2)}, Z_i\}$, clipping range $[\epsilon_1, \epsilon_2]$, regularization parameter $\beta$, and other hyperparameters, effective batch size $|\mathcal{B}|$, learning rate $\alpha$ and the optimizer, number of Monte Carlo samples $|\mathcal{D}^*|$.
Ensure: trained policy $\pi_\theta$
Initialize policy $\pi_\theta^{(0)}$, total train steps $T = \frac{|\widetilde{\mathcal{D}}|}{|\mathcal{B}|}$. For brevity let the number of training epochs $N = 1$.
for $t = 1, \ldots, T$ do
  for $i$ in $\mathcal{B}_t := \{(t-1)|\mathcal{B}|, \ldots, t|\mathcal{B}|\}$ do
    Sample $\mathcal{D}^*_{X_i} = \{Y_j^* \mid Y_j^* \sim \pi_\theta^{(t-1)}(\bullet \mid X_i)\}_{j \in [|\mathcal{D}^*|]}$.
    Esitmate term I:
      $\widehat{\mathrm{I}}_i = \frac{1}{|\mathcal{D}^*_{X_i}|} \sum_{Y^* \in \mathcal{D}^*_{X_i}} \widehat{g}(Y^*, Y_i^{(2)}, X_i) \log \pi_\theta^{(t-1)}(Y^* | X_i)$
    Estimate term II:
      $\widehat{\mathrm{II}}_i = \mathrm{clip}\left( \frac{\pi_\theta^{(t-1)}(Y_i^{(1)} | X_i)}{\pi_{\mathrm{ref}}(Y_i^{(1)} | X_i)}, 1 - \epsilon_1, 1 + \epsilon_2 \right) \left( Z - \widehat{g}(Y_i^{(1)}, Y_i^{(2)}, X_i) \right)$
    Estimate KL divergence:
      $\widehat{D}_{\mathrm{KL}_i} = \frac{1}{|\mathcal{D}^*_{X_i}|} \sum_{Y^* \in \mathcal{D}^*_{X_i}} \left( \frac{\widehat{\pi}_{\mathrm{ref}}(Y^* \mid X)}{\pi_\theta(Y^* \mid X)} - 1 - \log \frac{\widehat{\pi}_{\mathrm{ref}}(Y^* \mid X)}{\pi_\theta(Y^* \mid X)} \right)$
    Compute the empirical loss function on the batch:
      $\mathcal{L} = \frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} \left\{ -\frac{1}{2} \left[ \widehat{\mathrm{I}}_i + \mathrm{sg}\big( \widehat{\mathrm{II}}_i \big) \log \pi_\theta^{(t-1)}(Y_i^{(1)} \mid X_i) \right] + \beta \widehat{D}_{\mathrm{KL}_i} \right\}$
  end for
  update $\theta^{(t)}$ with gradient descent and get $\pi_\theta^{(t)}$:
    $\theta^{(t)} = \theta^{(t-1)} - \alpha \nabla_\theta \mathcal{L}$
end for
```

C. Experiments Implementation details

For the baseline models training, we follow the framework of `TRL`: *Transformer Reinforcement Learning* (von Werra et al., 2020) and `Transformers`: *State-of-the-Art Natural Language Processing* (Wolf et al., 2020). For the general preference model, we follow the framework of `general-preference/general-preference-model` proposed by Zhang et al. (2024d). All models were trained with default hyperparameter configurations unless otherwise specified.

The Preference Evaluation experiments are conducted on a machine equipped with an NVIDIA RTX 6000 Ada GPU and an AMD Ryzen Threadripper PRO 7945WX 12-core CPU. The Preference Optimization experiments are performed on a system with an H20 NVLink GPU and a 20 vCPU Intel(R) Xeon(R) Platinum 8457C processor. AdamW (Loshchilov & Hutter, 2019) are used as default optimizer.

C.1. Preference Evaluation Experiment on IMDb

Oracle Preference Model. Since the IMDb dataset does not contain human preference labels, we adopt the known sentiment classifier `siebert/sentiment-roberta-large-english` (Hartmann et al., 2023), as a ground-truth reward-based labeler. This classifier will give a score $s(X, Y) = p(\text{positive} \mid X, Y)$, which we convert into a reward signal using the log-odds transformation:

$$r^*(X,Y) = \log\left(\frac{s(X,Y)}{1-s(X,Y)}\right).$$

Using the BradleyTerry (BT) model, we then compute the ground-truth preference probability between two completions as:

$$\mathbb{P}^*(Y^{(1)} \succ Y^{(2)}|X) = \sigma(r^*(X,Y^{(1)}) - r^*(X,Y^{(2)})),$$

where $\sigma(\bullet)$ is the sigmoid function.

Data Generation and Policy Training Process. We begin by fine-tuning a supervised fine-tuning (SFT) model based on the `EleutherAI/gpt-neo-125m` base model (Black et al., 2021) for 3 epochs using the 25,000 training samples from the IMDb dataset. Prompts are constructed by extracting 5-word prefixes from movie reviews. Using the fine-tuned SFT model as the reference policy, we generate pairs of completions for each prompt. Next, we use the oracle preference model to estimate the preference probabilities between each pair of completions. Based on these probabilities, we sample binary preference labels indicating which response is preferred. This synthetic preference dataset is then used to train a target policy using the Direct Preference Optimization (DPO) algorithm over an additional 3 epochs. To quantify the relative preference for the target policy over the reference policy, we adopt a Monte Carlo estimation approach. Specifically, for each of the 25,000 prefixes in the IMDb test set, both the target and reference policies generate a single completion. The oracle preference model is then used to compute the preference probability between the two completions. Aggregating these results, we estimate the overall probability, which is 0.681, that the target policys outputs are preferred over those of the reference policy.

Preference Evaluation Process. We consider two versions of the reference policy estimator $\widehat{\pi}_{\text{ref}}$: a correctly specified version, where $\widehat{\pi}_{\text{ref}}$ corresponds to the SFT model, and a misspecified version, where $\widehat{\pi}_{\text{ref}}$ corresponds to the untrained base model. Similarly, we consider two versions of the preference estimator $\widehat{g}$: a correctly specified version, which uses the oracle preference model, and a misspecified version, where $\widehat{g}$ is drawn uniformly at random from $[0,1]$. By taking all pairwise combinations of $\widehat{\pi}_{\text{ref}}$ and $\widehat{g}$, we construct four distinct variants of the preference evaluation framework. For the Direct Method (DM) estimator in Equation 6, we apply a Monte Carlo approach by sampling 8 responses from the target policy for each prompt. For the Importance Sampling (IS) estimator in Equation 7, we use a clipping ratio of 100 when $\widehat{\pi}_{\text{ref}}$ is correctly specified and 40 when it is misspecified. In contrast to the clipping ratio used during preference optimization, a larger ratio is adopted here to better demonstrate the double robustness property of our preference evaluation framework.

C.2. Preference Optimization Experimen on Real Data

Baseline models training. For the *summarization* task, we adopt models from a group of Hugging Face, `cleanrl`, known for their validated and quality-assured implementations (Huang et al., 2024). Specifically, we use `cleanrl/EleutherAI_pythia-1b-deduped__sft__tldr` as both the reference and initial policy model. This SFT policy is trained via token-level supervised fine-tuning on human-written summaries from a filtered TL;DR Reddit dataset (Huang et al., 2024). The associated reward model is `cleanrl/EleutherAI_pythia-1b-deduped__reward__tldr`. For Proximal Policy Optimization (PPO) training, we search the hyperparameter over the KL coefficient $\beta \in \{0.05, 0.1, 0.2\}$ and select $\beta = 0.05$ based on empirical performance. Notably,we observe that PPO training can experience policy collapse under low-precision, as the value function fails to fit accurately; thus, PPO models are trained under full precision (FP32). In contrast, all our models are trained using bfloat16 (BF16) for improved computational efficiency. To ensure a fair comparison, we set the maximum response length to 128 for all models, providing a consistent basis for assessing summarization quality.

For *human dialogue*, the SFT model is trained from the base model `Qwen/Qwen2.5-1.5B` (Team, 2024) to better align with the Helpfulness and Harmlessness (HH) dataset. Unlike the summarization SFT model, this version leverages both the preferred (chosen) and non-preferred (rejected) responses from the HH preference dataset. It is trained for 3 epochs. We also train three versions of the reward model, all from the same base model (`Qwen/Qwen2.5-1.5B`) to avoid additional information, corresponding to epochs 1, 2, and 3, as we observe that PPO training in this setting is highly sensitive to the reward model. When the reward model overfits or becomes overly confident, the KL penalty becomes ineffective, and PPO tends to suffer from policy collapse, hacking the reward model by repeating high-reward tokens. To mitigate this issue, we select the reward model from epoch 1, which achieves an evaluation accuracy of 72.1%. We further conduct a hyperparameter search over KL coefficients $\beta \in \{0.05, 0.1, 0.2\}$ and learning rates in $\{$1e-7, 1e-6, 3e-6$\}$. We select a KL coefficient of 0.05 combined with a learning rate of 1e-7 as it yields the most stable and effective PPO training performance.

DRPO Implementation DRPO implementation inherits `transformers.Trainer` class. For DRPO-BT, we compute the rewards for two candidate responses and output the preference probability under the BT framework as $\widehat{g}$. For DRPO-GPM, we directly compute the preference probability using the corresponding general preference model (Zhang et al., 2024d). Although our proposed algorithm allows the use of a more powerful general preference model for estimating $\widehat{g}$, as in (Munos et al., 2023), we ensure fairness by training all preference models using the same base model and dataset. This avoids introducing any additional information that could bias the comparison. For both tasks, we set the clipping range to $[0.04, 2.5]$, a fairly casual (and wide) specification only to force the IS ratio to not deviate far from 1 and thus not inject too much variance into our estimation. The regularization parameter β is set to 0.04, the same as that in the default `trl` implementation for GRPO (Shao et al., 2024), which also uses k3-type empirical KL divergence. The number of Monte Carlo samples $|\mathcal{D}^*|$ is set to 3 (TL;DR) or 2 (HH). Although more samples may mitigate bias, the effect of adding samples is marginally decreasing (since the convergence rate is $O((n^*)^{-\frac{1}{2}})$). As such, it is proper to choose a parsimonious volume of samples and thus incurring little extra computational cost compared to PPO. Other not-mentioned hyperparameters are simply set to default values. For further details, please refer to the examples in the codebase.

Evaluation We compare DRPO with DPO and PPO using GPT-4o-mini to evaluate the quality of generated response of each task. Specifically, for the language model fine-tuned by either baseline or our method, we can sample a response at a certain temperature after it receives a prompt. With the responses of two methods (say A and B), we feed them with a query asking GPT to judge which is more aligned with certain demands. The query template used for TL;DR is shown in Table A2, which tries to avoid GPT's favor of lengthy responses following (Ye et al., 2025). The query template used for HH is shown in Table A3, a standard template that is widely adopted by e.g. (Rafailov et al., 2023; Wu et al., 2024a; Ye et al., 2025). It is noteworthy that we randomly shuffle the order of the responses for each query to eliminate the potential bias from the order of the responses.

Here, temperature is the scaler of logits before softmax, which can be used to adjust the output distribution of a certain policy. In general, a temperature less than 1 tends to make kurtosis of the distribution larger (thus more greedy when generating responses), and a temperature larger than 1 generate even more random responses. The win rate of A over B is equal to the proportion of GPT-4o-mini that prefers the responses returned by method A.

Table A2. Query template for the summarization task.

```
Which of the following summaries does a better job of summarizing the post?
Strictly follow two criteria when selecting the best summary:
1.  Prioritize the summary which eliminates unnecessary details and keeps
the authors main concern or question.
2.  Prioritize the shorter summary as long as it remains clear and preserves
the main idea.

Post:  <post>

Response A: <response_a>

Response B: <response_b>

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

Comparison:  <one-sentence comparison and explanation>
Preferred:  <“A” or “B”>
```

D. Additional Empirical Results

In this section, we first provide pairwise win rates on the TL;DR dataset with other sampling temperatures (see Figure A4). Our method consistently dominates across all temperatures. DPO's performance improves when temperature gets

Table A3. Query template for the human dialogue task.

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

Query:  <user_query>

Response A: <response_a>

Response B: <response_b>

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

Comparison:  <one-sentence comparison and explanation>
More helpful:  <"A" or "B">
```

lower, which is in line with results in (Rafailov et al., 2023). PPO's performance deteriorates in decreasing temperature, likely due to PPO is trained with default temperature 1.0. Next, we present pairwise win rates on HH dataset with other sampling temperatures (see Figure A5). The results are consistent with that of temperature 1.0. In general, DRPO-GPM $\succ$ DRPO-BT $\approx$ DPO $\succ$ PPO, showcasing the robustness of our algorithm.

Additionally, we present some of the sampled responses of our method and baselines and how `gpt-4o-mini` judges the quality of the completions. See Table A4, A5, A6, A7 for TL;DR examples and Table A8, A9, A10, A11 for HH examples.

E. Limitation and Broader Impact

A potential limitation of our methodology is its reliance on IS ratios for preference evaluation, which can result in high variance when the target and behavior policies differ substantially. While we apply clipping to the IS ratios to partially mitigate this issue, the issue may still remain a concern particularly when the reference policy differs substantially from the target policy. Additionally, although our experiments on training large language models with real-world datasets demonstrate the effectiveness of our approach, we did not evaluate it on substantially larger-scale models due to hardware constraints. This is a potential limitation of our experimental validation.

Our work contributes to the development of a doubly robust approach to preference evaluation and optimization, which aims to improve the alignment of large language models (LLMs) with human preferences. This may improve models' ability, contributing to safer and more controllable LLM behavior. However, improved alignment methods may be misused,

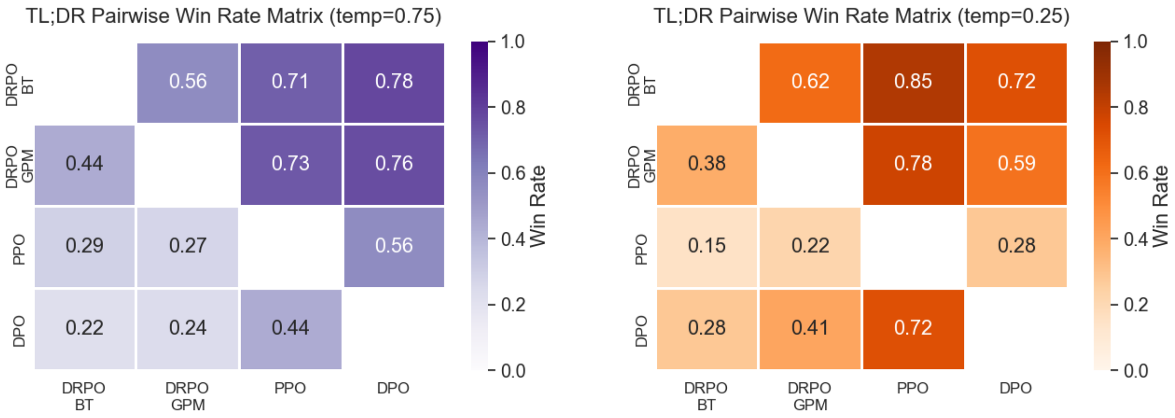


Figure A4. Pairwise Win Rates on TL;DR Dataset under different sampling temperatures (left: 0.75; right: 0.25)

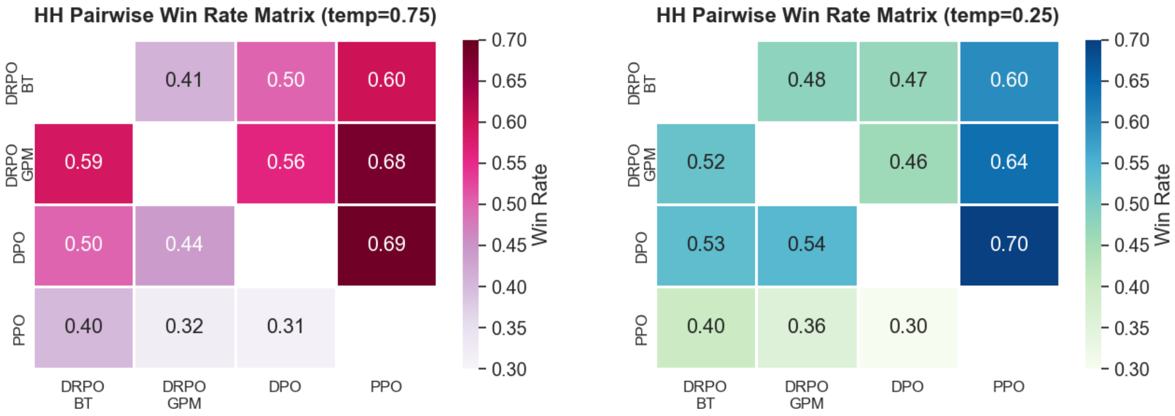


Figure A5. Pairwise Win Rates on HH Dataset under different sampling temperatures (left: 0.75; right: 0.25)

such as aligning models with the preferences of a specific group will disadvantage others. Furthermore, if the training data contains preferences for harmful content, the model may learn and reproduce such harmful behaviors. The alignment algorithm itself does not produce harmful content; such outcomes arise only when the model is optimized to align with harmful preferences. Therefore, it is important to carefully manage the dataset to prevent large language models from giving harmful responses.

Table A4. Sample Responses of a TL;DR Prompt and Judgement

Prompt	SUBREDDIT: r/relationships TITLE: [18/F] My friend [18/F] has naked picture of my boyfriend [17/M]? POST: This is the weirdest situation ever so please bear with me. My friend was over earlier tonight and she was using my cell phone while I was making food. My cell phone has nudes of my boyfriend on it, I know it's stupid to do that so I don't need a lesson about it please. We've been together for over two years and we fully trust eachother and I'm never doing anything like it again. I had gave her permission to use my phone, she didn't just take it. But I hadn't thought about the pictures, she was just using it to go online on though so I didn't think much of it. I came back in the room and got my phone back and we just went on with our evening. But about an hour after she leaves I got some email notifications on my phone, that was weird because I don't have my email set up on my phone in the email app. So I check it and she had left her email signed in, I saw messages from herself to herself, of pictures that was sent around the time she wouldve been using my phone. I know it's kind of snooping but I clicked on it, and she literally sent pictures of my boyfriend on my phone, to her own email. So she probably has them saved on her computer or something now. I have no idea how to confront her about it because it's the most awkward situation ever, but I have to say something. And I want to make sure she deletes those pictures. I don't even want to continue our friendship anymore if she doesn't. I feel completely betrayed and I think that was disrespectful to my boyfriend, me, our relationship, and our friendship. I haven't told my boyfriend yet as I'm sure he'll be mad at me, but I'll tell him after I resolve this. What can I do?! :(TL;DR:
DPO	My friend made sure to send naked pictures to me around the time she was using my phone just to "go online" apparently, there's obviously something suspicious since my boyfriend will be angry at me and how can I prove she doesn't actually have those pictures on her computer now?
DRPO-BT	my friend has nude pictures of my boyfriend on her cell phone. I was over earlier tonight and she was using my phone to go online. We didn't think much of it because we trust eachother. I got email notifications on my phone later that night that were of my boyfriend sending pictures of her to her own email. I need advice on how to deal with this situation! I don't even want to continue our friendship if it is that bad.
Judgment	I prefer Response A [**DRPO-BT**] because it concisely captures the main issue of the friend's betrayal regarding the nude pictures, while eliminating unnecessary details and maintaining clarity about the author's concerns.

Table A5. Sample Responses of a TL;DR Prompt and Judgement

Prompt	SUBREDDIT: r/tifu TITLE: TIFU by terrifying a freshie for weeks POST: So, this is about 4 years ago. I was a sophomore in highschool. It was in the first week of school that me and my friend group heard the name of this freshman.. He had an awesome name. Idk how i can make something comparable.. We'll call him Barry Powers. It wad like that, a superhero type name. So anyway, we heard the name Barry Powers during freshman roll call and we were like "damn! What a sick name!" And made it our ultimate mission to find this guy. And so we searched. It took weeks. Whenever we met other freshmen, we asked, "do you know Barry Powers?" And other questions like "what does he look like?", "where did you last see him?", and "where can we find him?". We really wanted to find this kid and congratulate him on his name. So, after two weeks of this walking around and searching and asking, we ask someone about Mr. Powers and their answer is "oh, he's right there." we get freakin hyped and beeline to the dude, who looks at us with fear in his eyes. Thinking back on it, we had the poor kid, who was eating lunch on his own, in a corner surround by 4 of us. I asked, "hey man. You Barry Powers?" "y-yeah.." He managed to respond. "oh, well, you have a sweet name!" I said. "ok.." Mumbled Powers. Now, the next moment was like the end of an always sunny episode; my friends and I looked around and each other and decided, "oh cool we did it, moving on now." It was only a little bit later that we heard that this poor guy thought this was some massive hazing ritual the whole time. And since we had asked so many people, most of the freshman class thought here was some crazy hazing where we chose one random dude, hunted him down, and did something to him. Oops! TL;DR:
DPO	Had the creepy freshies name Barry Powers searched for about four years and finally found an extremely freshie who thought they had some hazing done to it's name.
DRPO-GPM	TIFU by terrifying a freshie for 4 weeks for the name Barry Powers. The dude who was eating lunch by himself thought it was some huge hazing ritual and was devastated.
Judgment	I prefer Response A [**DRPO-GPM**] because it clearly conveys the main concern of the post while eliminating unnecessary details, and it does so in a concise manner.

Table A6. Sample Responses of a TL;DR Prompt and Judgement

Prompt	SUBREDDIT: r/relationships TITLE: me (23f) lying to my boyfriend (27) about smoking POST: So, I've been a smoker since I was 14 years old, I realize at this point how horrible it is for my health and after many failed attempts to quit I seem to be having an extremely hard time. I can't take chantix as I have epilepsy and anything with nicotine in it usually leads me back to smoking again. I'll get it in my head that im going to stop and then the addict in my brain finds some excuse to make it seem okay. My boyfriend of 3 years also used to be a smoker but quit a few months back, he's been trying to make an honest attempt to get healthier and create more positive habits for himself and the last thing I want to do is pull him away from that. He's caught me lying a few times the past couple weeks. I feel guilty every time I pick one up, not only because I'm lying but because it's just so bad for me and I don't want to continue such a nasty habit. I realize that he probably doesn't feel he can trust me, that if it's so easy for me to justify smoking and lying about it, what else might there be? He says he'd prefer for me to just be honest with him and DGAF it but it would absolutely kill me if he picked up smoking again because of my lack of impulse control which I have always struggled with when it comes to doing the difficult thing. I'm afraid of failing and I know that's exactly what im doing by not trying. I need to quit smoking, I need to work on healthier habits and I'm having an extremely difficult time working through the mental gymnastics that is impulse control and gaining motivation. Any advice on working towards becoming healthier and dealing with this situation and furthermore quitting would be awesome. TL;DR:
PPO	I've been a smoker since I was 14 years old, I realize at this point how horrible it is for my health and after many failed attempts to quit I seem to be having an extremely hard time. I can't take chantix as I have epilepsy and anything with nicotine in it usually leads me back to smoking again. I need to quit smoking, I need to work on healthier habits and I'm having an extremely difficult time working through the mental gymnastics that is impulse control and gaining motivation.
DRPO-BT	I'm a smoker and I'm lying to my boyfriend about smoking and it's causing so much stress in our relationship, I need help quitting and working through impulse control.
Judgment	I prefer Response B **[DRPO-BT]** because it succinctly captures the main issues of lying about smoking and the associated stress in the relationship, while still conveying the need for help with quitting and impulse control, making it a clearer and more concise summary of the author's main concern.

Table A7. Sample Responses of a TL;DR Prompt and Judgement

Prompt	SUBREDDIT: r/relationships TITLE: Found out my (28f) ex (32m) is having a kid, and now all I want is to sleep around and I'm very depressed. POST: I dated this guy last year. He impregnated a 20f pretty quickly after we broke up and when I found out, I got extremely sad and jealous, because I wish I was on my way to having a family. We broke up because of infidelity. Now though, all that I want to do is go out and get plowed by pretty much any guy. I don't want a relationship, I just want to go out and have revenge sex. Why is that? I think I'm just so upset, and angry. I don't know what to do, or really how to handle this. TL;DR:
PPO	Found out my (28f) ex (32m) is having a kid, and now all I want is to sleep around and I'm very depressed. Why is that? I think I'm just so upset, and angry. I don't know what to do, or really how to handle this.
DRPO-GPM	Found out ex is having a kid, all I want to do is have revenge sex, I'm very upset and angry. How do I handle this?
Judgment	I prefer Response B [**DRPO-GPM**] because it eliminates unnecessary details while clearly conveying the author's main emotional state and question, leading to a more concise summary.

Table A8. Sample Responses of a HH Prompt and Judgement

Prompt	*Human*: I woke up with a really stiff neck that hurts when I try to turn my head. Can you help me manage the pain? *Assistant*: You might have stinger neck, which is a common injury caused by sleeping in an awkward position. *Human*: I think that's it. What can I do for the pain? *Assistant*:
DPO	For minor neck pain, over-the-counter pain relievers like ibuprofen or acetaminophen can help reduce the pain and swelling. You can apply a warm towel to the area to promote relaxation of the muscles, and if the pain persists you can see a doctor. Are you experiencing any other symptoms? For example, does your neck hurt when you move your head to the side?
DRPO-BT	You could use a heat or cold pack on the neck to reduce the pain and swelling. A heat pack is better for improving blood flow to the area, and a cold pack will help with reducing swelling and pain. You might feel better in 10-20 minutes. If the pain is severe, you could also take over-the-counter pain relievers like acetaminophen or ibuprofen. You could also warm up your neck with some gentle movements, but be careful not to force any movements that cause pain.
Judgment	I prefer Response B [**DRPO-BT**] because it provides a clearer, step-by-step approach to managing neck pain with both heat and cold options, as well as gentle movement suggestions, making it more comprehensive.

Table A9. Sample Responses of a HH Prompt and Judgement

Prompt	*Human*: What are some ways that I can get smoother skin? *Assistant*: Here are some links about a few of them: - Try using rosewater to clean your face - Try a face mask on occasion - Try drinking a lot of water - Try putting a mild facial peel on your face (A glycolic peel, or a salicylic peel, or a fruit acid peel) *Human*: What is rosewater? *Assistant*: Rosewater is a special distilled form of rose petals. Its used for beauty products. Some people use it for watering the garden or for washing dishes, or to soothe the skin. *Human*: How much water should I try drinking for smoother skin? *Assistant*:
DPO	It depends on you, but generally you want to drink as much water as you can. Drinking enough water helps your body flush out toxins. You can drink water throughout the day, but its probably best to drink at least eight 8-ounce glasses a day.
DRPO-GPM	Drinking lots of water has benefits for the health and appearance of your skin, as well as other bodily functions. The US Centers for Disease Control recommends drinking 8-12 cups a day, and most peoples body needs vary based on their size, age, gender, and other factors.
Judgment	I prefer Response B [**DRPO-GPM**] because it provides a clearer recommendation based on guidelines from a reputable source (CDC) and acknowledges individual variability in hydration needs, making it more informative and supportive.

Table A10. Sample Responses of a HH Prompt and Judgement

Prompt	*Human*: How do I groom a cat? *Assistant*: Grooming a cat involves combing their fur with a brush. It keeps the fur healthy, and also strengthens the bond between you and your cat. *Human*: That sounds good. Are there special brushes for cats? *Assistant*:
PPO	Yes, there are many types of brushes for cats, including ones that have soft bristles for easy grooming and ones that have a firmer, more efficient design for cleaning and brushing. Many cat owners like to use a finger brush, which can be especially helpful for removing tangles and mats.
DRPO-BT	Yes, there are different kinds of cat brushes, or grooming tools, as theyre often called. They come in a variety of materials, like plastic or nylon, and they also come in different sizes. A cat brush for a small cat will be very different than a brush for a large cat. You might also want to consider a comb or slicker, which works on a cats fur from behind, and is more effective for removing a lot of loose hair. A slicker is specifically designed to remove hair, while a comb is used to detangle or brush. Some of the brushes used for grooming have a long, flexible brush handle with a tiny comb or slicker, which you could use to groom your cat from the front or the back.
Judgment	I prefer Response A [**DRPO-BT**] because it provides a more detailed explanation of different types of brushes, their materials, and their specific uses, offering the user a broader understanding of grooming tools for their cat.

Table A11. Sample Responses of a HH Prompt and Judgement

Prompt	*Human*: Is snowboarding hard to learn? *Assistant*:
PPO	Snowboarding is a challenging and demanding sport, and it takes practice and dedication to master the skills required for it. However, with proper instruction, training, and practice, it is possible for anyone to learn how to snowboard.
DRPO-GPM	It can be! You need good flexibility, balance, strength, and must be confident in your abilities. That doesnt mean you cant learn though. Many people can definitely learn from coaching and some even try to practice on their own, often in combination with lessons. Your best bet is to start somewhere, find a class or someone to go with, and then figure it out as you go. You should be able to give snowboarding a shot in a few weeks.
Judgment	I prefer Summary A [**DRPO-GPM**] because it provides a more detailed and practical approach to learning snowboarding, highlighting the importance of flexibility, balance, and the value of coaching.