Doubly Robust Alignment for Large Language Models: Technical Appendices

Anonymous Author(s)
Affiliation
Address
email

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.

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., Radon–Nikodym 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 S1. *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., 1, 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 S1. 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

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

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. [2] and Theorem 3.5 in Tsiatis [1]. See also the proof of Theorem 1 in [3]. 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.

$$\text{(i)}\ \mathbb{E}\left\{\psi(W)\frac{\partial}{\partial\gamma}\log l(W;\gamma_0,b_0,\eta_0)\right\} = \frac{\partial}{\partial\gamma}p^*(\pi)|_{\gamma=\gamma_0},$$

$$\text{(ii)}\ \mathbb{E}\left\{\psi(W)\frac{\partial}{\partial b}\log l(W;\gamma_0,b_0,\eta_0)\right\} = \frac{\partial}{\partial b}p^*(\pi)|_{b=b_0},$$

$$\text{(iii)}\ \mathbb{E}\left\{\psi(W)\frac{\partial}{\partial\eta}\log l(W;\gamma_0,b_0,\eta_0)\right\} = \frac{\partial}{\partial\eta}p^*(\pi)_{\eta=\eta_0}.$$

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 (S.1) that

$$\begin{aligned} \log l(w;\gamma,b,\eta) &= \log f_\gamma(x) + \log\pi_b(y_1|x) + \log\pi_b(y_2|x) \\ &\quad + z\log g_\eta(y_1,y_2,x) + (1-z)\log(1-g_\eta(y_1,y_2,x)). \end{aligned} \tag{S.2}$$

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 $\mathrm{Ber}(p)$ denote the Bernoulli distribution with success probability p. The left-hand-side (LHS) of equation (i) can be represented by

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

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

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

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

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

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

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

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

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

On the other hand, since

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

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

$$
\begin{aligned}
& \frac{1}{2}\sum_{x,y_1,y_2}\left(\frac{\pi(y_1|x)}{\pi_{\mathrm{ref}}(y_1|x)}-\frac{\pi(y_2|x)}{\pi_{\mathrm{ref}}(y_2|x)}\right)\pi_{\mathrm{ref}}(y_1|x)\pi_{\mathrm{ref}}(y_2|x)\frac{\partial}{\partial\eta}g_\eta(y_1,y_2,x)f_X(x) \\
= \;& \frac{1}{2}\sum_{x,y_1,y_2}(\pi(y_1|x)\pi_{\mathrm{ref}}(y_2|x)-\pi(y_2|x)\pi_{\mathrm{ref}}(y_1|x))\frac{\partial}{\partial\eta}g_\eta(y_1,y_2,x)f_X(x) \\
= \;& \sum_{x,y_1,y_2}\pi(y_1|x)\pi_{\mathrm{ref}}(y_2|x)\frac{\partial}{\partial\eta}g_\eta(y_1,y_2,x)f_X(x) \\
= \;& \frac{\partial}{\partial\eta}p^*(\pi)
\end{aligned}
\tag{S.3}
$$

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 [3]. Based on the likelihood function in equation (S.2), 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

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

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

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

where

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

From Lemma S1, 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 $\text{MSE}(\widehat{p}_{\text{DR}})$:

$$
\text{MSE}(\widehat{p}_{\text{DR}}(\pi)) = \text{Var}(\mathbb{E}_n\psi(w;\pi,\pi_{\text{ref}},g^*)+\mathrm{I}+\mathrm{II}+\mathrm{III})+(\mathbb{E}[\mathrm{III}])^2 \tag{S.4}
$$

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

$$
\begin{aligned}
\text{Var}(\mathbb{E}_n\psi(w;\pi,\pi_{\text{ref}},g^*) = \frac{1}{n}\text{Var}(\psi(w;\pi,\pi_{\text{ref}},g^*)) &= O\Big(\frac{1}{n}\mathbb{E}\frac{\pi^2(Y|X)}{\pi_{\text{ref}}^2(Y|X)}\Big)\\
&= O\Big(\frac{1}{n}\sum_y\frac{\pi^2(y|X)}{\pi_{\text{ref}}(y|X)}\Big) = O\left(\frac{1}{n\epsilon}\right).
\end{aligned} \tag{S.5}
$$

Moreover, we have

$$
\begin{aligned}
\mathbb{E}\mathrm{I}^2 =&\frac{1}{2n}\mathbb{E}\left\{(Z-g^*(X,Y^{(1)},Y^{(2)})^2\left[\frac{\pi(Y|X)}{\widehat{\pi}_{\text{ref}}(Y|X)}-\frac{\pi(Y|X)}{\pi_{\text{ref}}(Y|X)}\right]^2\right\}\\
\leq&\frac{1}{2n}\mathbb{E}\left\{\frac{\pi^2(Y|X)}{\pi_{\text{ref}}^2(Y|X)}\left[\frac{\pi_{\text{ref}}(Y|X)}{\widehat{\pi}_{\text{ref}}(Y|X)}-1\right]^2\right\}\\
=&\frac{1}{2n}\mathbb{E}_X\left\{\sum_y\frac{\pi^2(y|X)}{\pi_{\text{ref}}(y|X)}\left[\frac{\pi_{\text{ref}}(y|X)}{\widehat{\pi}_{\text{ref}}(y|X)}-1\right]^2\right\}\\
=&O\left(\frac{1}{n\epsilon}\left\|\frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}}-1\right\|^2\right).
\end{aligned} \tag{S.6}
$$

Follow a similar argument, we obtain

$$
\mathbb{E}\mathrm{II}^2 \;=\; O\left(\frac{1}{n\epsilon}\|\widehat{g}-g^*\|^2\right),\qquad \mathbb{E}\mathrm{III}^2 = O\left(\frac{1}{n\epsilon}\left\|\frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}}-1\right\|^2\right). \tag{S.7}
$$

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

$$
\begin{aligned}
\text{Cov}\left(\mathbb{E}_n\psi(w;\pi,\pi_{\text{ref}},g^*),\mathrm{I}+\mathrm{III}\right) =&O\left(\frac{1}{n\epsilon}\left\|\frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}}-1\right\|\right),\\
\text{Cov}\left(\mathbb{E}_n\psi(w;\pi,\pi_{\text{ref}},g^*),\mathrm{II}\right) =&O\left(\frac{1}{n\epsilon}\|\widehat{g}-g^*\|\right),\\
\text{Cov}\left(\mathrm{I}+\mathrm{III},\mathrm{II}\right) =&O\left(\frac{1}{n\epsilon}\|\widehat{g}-g^*\|\cdot\left\|\frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}}-1\right\|\right).
\end{aligned} \tag{S.8}
$$

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 (S.8). Combining equations (S.5), (S.6),(S.7) and (S.8) yields

$$\text{Var}(\mathbb{E}_n\psi(w;\pi,\pi_{\text{ref}},g^*) + \text{I} + \text{II} + \text{III}) = \text{SEB} + O\left(\frac{1}{n\epsilon}\|\widehat{g} - g^*\|\right) + O\left(\frac{1}{n\epsilon}\left\|\frac{\widehat{\pi}_{\text{ref}}}{\pi_{\text{ref}}} - 1\right\|\right). \quad \text{(S.9)}$$

Finally, using Cauchy inequality again, we obtain that

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

Combining (S.4) and (S.10), we obtain

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

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 $\mathbb{E}_{X\sim\mathcal{D}} D_{\text{KL}}[\pi_1(\bullet \mid X) \,\|\, \pi_2(\bullet \mid X)]$ as $\text{KL}(\pi_1\|\pi_2)$. Since $\widehat{\pi}$ is a maximizer of $\widehat{p}_{\text{DR}}(\pi) - \beta\text{KL}(\pi\|\widehat{\pi}_{\text{ref}})$, we have

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

It directly follows that

$$\begin{aligned}
& p^*(\pi^*) - p^*(\widehat{\pi}) \\
\leq\; & p^*(\pi^*) - \widehat{p}_{\text{DR}}(\pi^*) + \widehat{p}_{\text{DR}}(\widehat{\pi}) - p^*(\widehat{\pi}) + \beta(\text{KL}(\pi^*\|\widehat{\pi}_{\text{ref}}) - \text{KL}(\widehat{\pi}\|\widehat{\pi}_{\text{ref}})) \\
\leq\; & \mathbb{E}\left|p^*(\pi^*) - \widehat{p}_{\text{DR}}(\pi^*)\right| + \mathbb{E}\left|\widehat{p}_{\text{DR}}(\widehat{\pi}) - p^*(\widehat{\pi})\right| + O\left(\beta\log^{-1}\epsilon\right) \\
\leq\; & 2\mathbb{E}\sup_{\pi\in\Pi}\left|p^*(\pi) - \widehat{p}_{\text{DR}}(\pi)\right| + O\left(\beta\log^{-1}\epsilon\right), \qquad \text{(S.10)}
\end{aligned}$$

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

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

$$\sup_{\pi\in\Pi}\left|\mathbb{E}[p^*(\pi) - \widehat{p}_{\text{DR}}(\pi)]\right| = \mathbb{E}|\mathbb{E}_n\psi(w;\pi,\pi_{\text{ref}},g^*) - p^*(\pi)| + O\left(\frac{1}{\epsilon}\|\widehat{g} - g^*\| \cdot \|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\|\right). \quad \text{(S.11)}$$

It remains to upper bound the empirical process term $\mathbb{E}\sup_{\pi\in\Pi}|\widehat{p}_{\text{DR}}(\pi) - \mathbb{E}\widehat{p}_{\text{DR}}(\pi)|$. Toward that end, we employ Corollary 5.1 in [4]. 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 [4] that the function class $\mathcal{F} = \{\psi(\bullet,\pi,\widehat{\pi}_{\text{ref}},\widehat{g})|\pi\in\Pi\}$ 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 [4] 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 (S.10), (S.11) and (S.12), 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., 5, 6]. Therefore,

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

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 $J(\pi^*)-J(\widehat{\pi})=O(\mathrm{Reg}(\widehat{\pi}))$ 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

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

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:

$$\begin{aligned}J(\pi^*)-J(\widehat{\pi}) &\leq \mathbb{E}\{[J(\pi^*)-l(\pi^*)]+[l(\widetilde{\pi})-l_n(\widetilde{\pi})]+[l_n(\widehat{\pi})-l(\widehat{\pi})]+[l(\widehat{\pi})-J(\widehat{\pi})]\}\\ &\leq \mathbb{E}\{[J(\pi^*)-l(\pi^*)]\}+\mathbb{E}\{[l(\widehat{\pi})-J(\widehat{\pi})]\}+2\mathbb{E}\sup_{\pi\in\Pi}\{|l(\pi)-l_n(\pi)|\}\end{aligned}\tag{S.12}$$

For the first term, we have

$$\begin{aligned}\mathbb{E}\{|J(\pi^*)-l(\pi^*)|\} &= \mathbb{E}_{y\sim\pi^*}|\widehat{r}(y,X)-r^*(y,X)|+\beta\mathrm{KL}(\pi^*\|\pi_{\mathrm{ref}})\\ &= \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}\left[\frac{\pi^*(y|X)}{\pi_{\mathrm{ref}}(y|X)}|\widehat{r}(y,X)-r^*(y,X)|\right]+O(\beta\log^{-1}\epsilon)\\ &= O\left(\frac{1}{\sqrt{\epsilon}}\|\widehat{r}-r^*\|\right)+O(\beta\log^{-1}\epsilon),\end{aligned}\tag{S.13}$$

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 $\mathcal{F} = \left\{\sum_y \widehat{r}(y, X)\pi(y|X)\big|\pi \in \Pi\right\}$ is bounded by a constant. Using similar arguments to the proof of Theorem 5, we can employ Corollary 5.1 in [4] to show that

$$\mathbb{E}\sup_{\pi\in\Pi}\{|l(\pi) - l_n(\pi)|\} = O\left(\frac{v}{n} + \sqrt{\frac{v}{n}}\right) + O(\beta \log^{-1}\epsilon). \tag{S.14}$$

Combining equations (S.12), (S.13) and (S.14), 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 [7]. We further assume

Assumption S.1 (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 S.2 (Coverage)**.** Both π_{ref} and $\widehat{\pi}_{\text{ref}}$ are lower bounded by some constant $\epsilon > 0$.

Assumption S.3 (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 S.1 and S.2 differ from those in the main text. Specifically, Assumption S.1 imposes the realizability assumption on the oracle reward rather than the optimal policy whereas Assumption S.2 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

$$r^\pi(y, x) = \beta\log\left(\frac{\pi(y|x)}{\widehat{\pi}_{\text{ref}}(y|x)}\right) + \beta Z(x) \tag{S.15}$$

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 (S.15) 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

$$\begin{aligned}
& \mathbb{E}_n l(\widetilde{\pi}) - \mathbb{E}_n l(\widehat{\pi}) - \mathbb{E} l(\widetilde{\pi}) + \mathbb{E} l(\widehat{\pi}) \\
\leq\; & \mathbb{E} l(\widehat{\pi}) - \mathbb{E} l(\widetilde{\pi}) \\
\leq\; & \mathbb{E} l(\widehat{\pi}) - \mathbb{E} l^*(\widetilde{\pi}) + \mathbb{E} l^*(\widetilde{\pi}) - \mathbb{E} l(\widetilde{\pi}) \\
\leq\; & -C_1 \mathbb{E}\,\|\widehat{r}(y_1, x) - \widehat{r}(y_2, x) - r^*(y_1, x) + r^*(y_2, x)\|_2^2 + \beta^2 C_2 \mathbb{E}\left(\log\frac{\widehat{\pi}_{\text{ref}}(Y^{(1)}|X)}{\pi_{\text{ref}}(Y^{(1)}|X)}\right)^2 \\
\leq\; & -C_1\sigma^2 + \beta^2 C_2\|\widehat{\pi}_{\text{ref}}/\pi_{\text{ref}} - 1\|^2,
\end{aligned} \tag{S.16}$$

where $\sigma^2 = \mathbb{E}\,\|\widehat{r}(y_1, x) - \widehat{r}(y_2, x) - r^*(y_1, x) + r^*(y_2, x)\|_2^2$, 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 [4], using similar arguments to the proof of Theorem 5 and PPO-based algorithms, we have

$$\begin{aligned}
\mathbb{E}_n l(\widetilde{\pi}) - \mathbb{E}_n l(\widehat{\pi}) - \mathbb{E} l(\widetilde{\pi}) + \mathbb{E} l(\widehat{\pi}) \;&\leq\; 2\mathbb{E}\sup_{\pi\in\Pi}|l(\pi) - \mathbb{E} l_n(\pi)| \\
&\leq\; O(\sigma\sqrt{\frac{v}{n}} + \frac{v}{n}).
\end{aligned} \tag{S.17}$$

This together with equation (S.16) 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

$$\sigma = O\left(\sqrt{\frac{v}{n}} + \beta\|\pi_{\text{ref}} - \widehat{\pi}_{\text{ref}}\|_2\right). \tag{S.18}$$

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

$$\begin{aligned}
&J(\pi^*) - J(\widehat{\pi}) \\
=&\mathbb{E}[\mathbb{E}_{y\sim\pi^*}r^*(y,X) - \mathbb{E}_{y\sim\widehat{\pi}}r^*(y,X)] \\
=&\mathbb{E}(\mathbb{E}_{y\sim\pi^*}r^*(y,X) - \mathbb{E}_{y\sim\widehat{\pi}^*}r^*(y,X)) + \mathbb{E}(\mathbb{E}_{y\sim\widehat{\pi}^*}r^*(y,X) - \mathbb{E}_{y\sim\widehat{\pi}}r^*(y,X)),
\end{aligned} \tag{S.19}$$

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

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

where the last equality is due to that under Assumptions S.2 and S.3, 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

$$\mathbb{E}\sum_y |\widehat{\pi}(y|X) - \widehat{\pi}^*(y|X)| \leq \frac{1}{\beta}\mathbb{E}\max_y|\widehat{r}(y,X) - r^*(y,X)| \leq \frac{1}{\beta\sqrt{\epsilon}}\|\widehat{r} - r^*\|_2, \tag{S.20}$$

where the last inequality follows from the fact that

$$\begin{aligned}
\|\widehat{r} - r^*\|_2 &= \mathbb{E}\{(\widehat{r} - r^*)^2\}^{1/2} \\
&= \mathbb{E}\left\{\sum_y \pi_{\text{ref}}(y|X)(\widehat{r}(y|X) - r^*(y|X))^2\right\}^{1/2} \\
&\geq \sqrt{\epsilon}\mathbb{E}\left\{\sum_y(\widehat{r}(y|X) - r^*(y|X))^2\right\}^{1/2} \\
&\geq \sqrt{\epsilon}\max_y|\widehat{r}(y,X) - r^*(y|X)|.
\end{aligned} \tag{S.21}$$

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

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

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 (S.20) using this difference. Consequently, we obtain that $\sigma^2 \geq \epsilon \|\widehat{r} - r^*\|^2$ and hence

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

Combining this together with equations (S.18) and (S.19), we obtain that the regret is upper bounded by

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

The proof is hence completed.

B DRPO Algorithm Details and Practical Implementation

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

$$\mathcal{J}(\pi_\theta; \widehat{\pi}_{\mathrm{ref}}, \widehat{g}_\eta, \mathcal{D}) = \widehat{p}_{\mathrm{DR}}(\pi) - \beta\mathbb{E}_{X\sim\mathcal{D}} D_{\mathrm{KL}}[\pi(\bullet \mid X) \,\|\, \widehat{\pi}_{\mathrm{ref}}(\bullet \mid X)]. \tag{S.22}$$

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

$$\begin{aligned}\nabla_\theta \mathcal{J}(\pi_\theta) \quad = \quad & \frac{1}{2}\mathbb{E}_{X,Y^{(1)},Y^{(2)}\sim\mathcal{D}}\left\{\sum_{a=1}^{2}\mathbb{E}_{y\sim\pi_\theta(\bullet|X)}\left[\widehat{g}(X,y,Y^{(a)})\nabla_\theta \log\pi_\theta(y|X)\right]\right.\\ & \left. + \sum_{a=1}^{2}(-1)^{a-1}\frac{\nabla_\theta\pi_\theta(Y^{(a)}|X)}{\widehat{\pi}_{\mathrm{ref}}(Y^{(a)}|X)}\left(Z - \widehat{g}(X,Y^{(1)},Y^{(2)})\right)\right\}\\ & -\beta\nabla_\theta D_{\mathrm{KL}}[\pi_\theta(\bullet \mid X) \,\|\, \widehat{\pi}_{\mathrm{ref}}(\bullet \mid X)]\end{aligned} \tag{S.23}$$

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 S.23. 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 [8].

$$\begin{aligned}\mathcal{L}_{\mathrm{DRPO}} \quad = \quad & -\frac{1}{2}\mathbb{E}_{X,Y^{(1)},Y^{(2)}\sim\mathcal{D}}\left\{\mathbb{E}_{Y^*\sim\mathcal{D}_X^*}\left[\sum_{a=1}^{2}\widehat{g}(Y^*,Y^{(a)},X)\log\pi_\theta(Y^*|X)\right]\right.\\ & \left. + \sum_{a=1}^{2}(-1)^{a-1}\frac{\pi_\theta(Y^{(a)}|X)}{\pi_{\mathrm{ref}}(Y^{(a)}|X)}\left(Z - \widehat{g}(X,Y^{(1)},Y^{(2)})\right)\right\}\\ & +\beta\mathbb{E}_{Y^*\sim\mathcal{D}_X^*,X\sim\mathcal{D}}\left[\frac{\widehat{\pi}_{\mathrm{ref}}(Y^* \mid X)}{\pi_\theta(Y^* \mid X)} - 1 - \log\frac{\widehat{\pi}_{\mathrm{ref}}(Y^* \mid X)}{\pi_\theta(Y^* \mid X)}\right]\end{aligned} \tag{S.24}$$

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 (S.24). 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:

$$
\begin{aligned}
\mathcal{L}_{\text{DRPO}} \quad = \quad & -\frac{1}{2}\mathbb{E}_{X,Y^{(1)},Y^{(2)}\sim\widetilde{\mathcal{D}}}\Bigg\{ \underbrace{\mathbb{E}_{Y^*\sim\mathcal{D}^*_X}\left[\widehat{g}(Y^*,Y^{(2)},X)\log\pi_\theta(Y^*|X)\right]}_{\text{term I}} \\
& + \text{sg}\Bigg(\underbrace{\text{clip}\Big(\frac{\pi_\theta(Y^{(1)}|X)}{\pi_{\text{ref}}(Y^{(1)}|X)}, 1-\epsilon_1, 1+\epsilon_2\Big)\big(Z-\widehat{g}(Y^{(1)},Y^{(2)},X)\big)}_{\text{term II}}\Bigg)\log\pi_\theta(Y^{(1)}\mid X)\Bigg\} \\
& + \beta\mathbb{E}_{Y^*\sim\mathcal{D}^*_X, X\sim\widetilde{\mathcal{D}}}\left[\frac{\widehat{\pi}_{\text{ref}}(Y^*\mid X)}{\pi_\theta(Y^*\mid X)} - 1 - \log\frac{\widehat{\pi}_{\text{ref}}(Y^*\mid X)}{\pi_\theta(Y^*\mid X)}\right]
\end{aligned}
\tag{S.25}
$$

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 $\widehat{\pi}_{\text{ref}}$, 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 β, and other hyperparameters, effective batch size $|\mathcal{B}|$, learning rate α and the optimizer, number of Monte Carlo samples $|\mathcal{D}^*|$.

Ensure: trained policy π_θ

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:

$$\widehat{\text{I}}_i = \frac{1}{|\mathcal{D}^*_{X_i}|}\sum_{Y^*\in\mathcal{D}^*_{X_i}} \widehat{g}(Y^*, Y_i^{(2)}, X_i)\log\pi_\theta^{(t-1)}(Y^*|X_i)$$

    Estimate term II:

$$\widehat{\text{II}}_i = \text{clip}\left(\frac{\pi_\theta^{(t-1)}(Y_i^{(1)}|X_i)}{\pi_{\text{ref}}(Y_i^{(1)}|X_i)}, 1-\epsilon_1, 1+\epsilon_2\right)\big(Z - \widehat{g}(Y_i^{(1)}, Y_i^{(2)}, X_i)\big)$$

    Estimate KL divergence:

$$\widehat{D}_{\text{KL}_i} = \frac{1}{|\mathcal{D}^*_{X_i}|}\sum_{Y^*\in\mathcal{D}^*_{X_i}}\left(\frac{\widehat{\pi}_{\text{ref}}(Y^*\mid X)}{\pi_\theta(Y^*\mid X)} - 1 - \log\frac{\widehat{\pi}_{\text{ref}}(Y^*\mid X)}{\pi_\theta(Y^*\mid X)}\right)$$

    Compute the empirical loss function on the batch:

$$\mathcal{L} = \frac{1}{|\mathcal{B}_t|}\sum_{i\in\mathcal{B}_t}\left\{-\frac{1}{2}\left[\widehat{\text{I}}_i + \text{sg}\big(\,\widehat{\text{II}}_i\,\big)\log\pi_\theta^{(t-1)}(Y_i^{(1)}\mid X_i)\right] + \beta\widehat{D}_{\text{KL}_i}\right\}$$

  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* [9] and `Transformers`: *State-of-the-Art Natural Language Processing* [10]. For the general preference model, we follow the framework of `general-preference/general-preference-model` proposed by Zhang et al. [11]. 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 [12] 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` [13], 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 Bradley–Terry (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 [14] 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 policy's 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 [15]. 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 [15]. 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` [16] 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 $\{1\text{e-}7, 1\text{e-}6, 3\text{e-}6\}$. 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 [11]. Although our proposed algorithm allows the use of a more powerful general preference model for estimating $\widehat{g}$, as in [17], 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

Table S1: 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 author's 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’’>
```

Table S2: 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’’>
```

to 0.04, the same as that in the default `trl` implementation for GRPO [8], 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 S1, which tries to avoid GPT's favor of lengthy responses following [18]. The query template used for HH is shown in Table S2, a standard template that is widely adopted by e.g. [7, 19, 18]. 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.

D Additional Empirical Results

In this section, we first provide pairwise win rates on the TL;DR dataset with other sampling temperatures (see Figure S1). Our method consistently dominates across all temperatures. DPO's performance improves when temperature gets lower, which is in line with results in [7]. 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 S2). 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 S3, S4, S5, S6 for TL;DR examples and Table S7, S8, S9, S10 for HH examples.

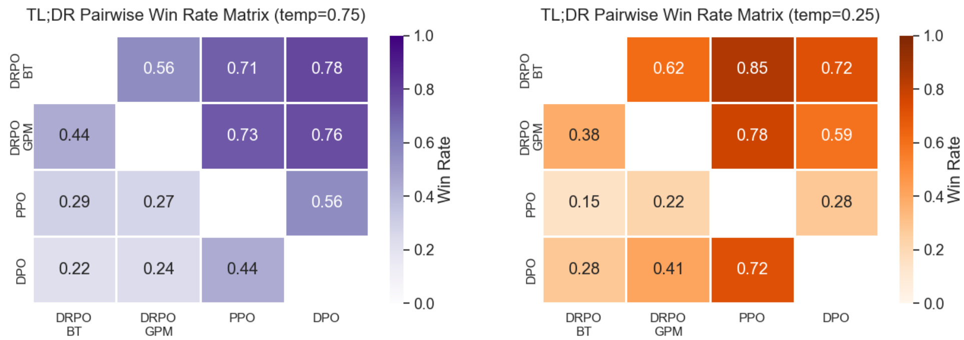

Figure S1: Pairwise Win Rates on TL;DR Dataset under different sampling temperatures (left: 0.75; right: 0.25)

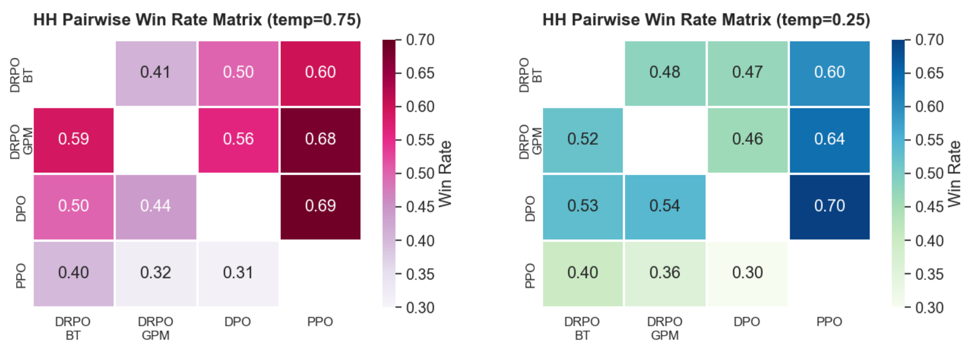

Figure S2: Pairwise Win Rates on HH Dataset under different sampling temperatures (left: 0.75; right: 0.25)

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, 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 S3: 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 S4: 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 S5: 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 S6: 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 S7: 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 S8: 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. It's 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 it's 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 people's 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 S9: 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 they're 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 cat's 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 S10: 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 doesn't mean you can't 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.

References

[1] Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.

[2] Peter J. Bickel, Chris A.J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York, 1998. ISBN 978-0-387-98473-5. URL https://link.springer.com/book/9780387984735.

[3] Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21:1–63, 2020.

[4] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, pages 1564–1597, 2014.

[5] Aaron K Han. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics*, 35(2-3):303–316, 1987.

[6] Robert P Sherman. The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society*, pages 123–137, 1993.

[7] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[8] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[9] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

[10] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[11] Yifan Zhang, Ge Zhang, Yue Wu, Kangping Xu, and Quanquan Gu. General preference modeling with preference representations for aligning language models. *arXiv preprint arXiv:2410.02197*, 2024.

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

[13] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 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.

[14] Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 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.

[15] Shengyi Huang, Michael Noukhovitch, Arian Hosseini, Kashif Rasul, Weixun Wang, and Lewis Tunstall. The n+ implementation details of rlhf with ppo: A case study on tl; dr summarization. *arXiv preprint arXiv:2403.17031*, 2024.

[16] Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

[17] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 18, 2023.

[18] Kai Ye, Hongyi Zhou, Jin Zhu, Francesco Quinzan, and Chengchun Shi. Robust reinforcement learning from human feedback for large language models fine-tuning. *arXiv preprint arXiv:2504.03784*, 2025.

[19] Tianhao Wu, Banghua Zhu, Ruoyu Zhang, Zhaojin Wen, Kannan Ramchandran, and Jiantao Jiao. Pairwise proximal policy optimization: Language model alignment with comparative rl. In *First Conference on Language Modeling*, 2024.