OTROPE: Optimal Transport-based Robust Off-policy Evaluation for Large Language Models

Liner Xiang linerx1@uci.edu
Wenbo Zhang wenbz13@uci.edu
Hengrui Cai* hengrc1@uci.edu
Department of Statistics, University of California, Irvine

Abstract

Reliable evaluation of large language models (LLMs) is essential for their development and deployment, yet is often costly, risky, and difficult to perform safely online. We study *off-policy evaluation* for LLMs, where limited human-labeled data from a behavior model are used to evaluate a newer target LLM. This setting is challenging because labels are scarce, behavior–target distribution shift is common, and response likelihoods are often unavailable for black-box LLMs. We propose the **O**ptimal **T**ransport-based **R**obust **O**ff-**P**olicy **E**valuation (**OTROPE**), a likelihood-free evaluation that performs *distributional correction* in a *semantic space* via optimal transport to align labeled behavior-policy samples with unlabeled target-policy samples. OTROPE combines corrected human-labeled residuals with proxy predictors, yielding a *doubly robust*-style evaluation without behavior-policy modeling or density-ratio estimation. We theoretically characterize why baseline evaluators fail under LLM distribution shift, and establish consistency and convergence rates for OTROPE when either the reweighted behavior distribution or the proxy predictor converges. Experiments on synthetic and real LLM evaluation tasks show that OTROPE consistently outperforms baselines while enabling ensembles of weaker LLM evaluators to approach and sometimes surpass stronger evaluators. Code is available at `https://github.com/LinerXiang/OTROPE`.

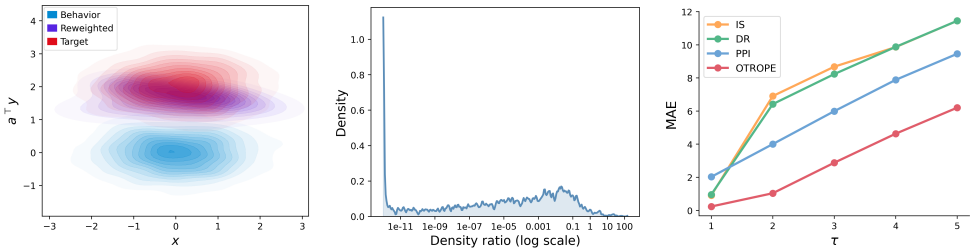
1 Introduction

As large language models (LLMs) continue to advance rapidly, reliable evaluation of their capabilities and behaviors has become increasingly important for their development and deployment. In many practical settings, evaluating an LLM can be viewed as evaluating the policy induced by the language model, which maps user prompts and contexts to generated responses (Bhargava et al., 2024; Huang et al., 2025). Although online evaluation provides direct evidence of model performance, it is often costly, risky, and difficult to perform safely. Offline evaluation using logged data is thus essential for efficiently and safely assessing the performance of LLMs. This naturally leads to *off-policy evaluation* (OPE) (Sutton et al., 1998), a central problem in reinforcement learning and decision-making that estimates the performance of a target policy from data collected under a different behavior policy. For LLMs, OPE provides a principled way to evaluate language models using historical prompts, responses, and feedback, thereby avoiding the cost and risk of direct online experimentation.

Offline evaluation of LLM policies faces three critical challenges. **First, ground-truth labels are often limited.** Evaluating LLMs requires reliable human feedback, but human evaluation is expensive and time-consuming, yielding only limited labeled data. As a scalable alternative, LLMs are increasingly used as proxy evaluators (Zheng et al., 2023; Yang et al., 2023; Fu et al., 2024); yet such proxies are not always trustworthy (Chen et al., 2024a; Ye et al., 2024; Schroeder and Wood-Doughty, 2024), making calibration or correction with *limited human labels* necessary. **Second, behavior and target policies often differ.** As LLMs evolve rapidly, human annotations

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) (b) (c)

Figure 1: (a) Density plots for behavior–target distribution shift, where the reweighted distribution by OTROPE closely matches the target one. (b) Density curve of behavior–target density ratios on labeled samples, showing a distribution concentrated near zero. (c) Mean absolute error (MAE) of different estimators with distributional shift level τ. OTROPE consistently achieves the lowest MAE.

typically collected for earlier models may lag behind the newer models being evaluated, creating behavior–target *distribution shift* (see Figure 1a). **Third, such shift is difficult to quantify in LLM settings.** Response likelihoods are often unavailable as users typically interact with LLMs via APIs or black-box systems. Estimating behavior policies from additional data can introduce policy misspecification and substantial computational cost. Even when such policies are estimated, sequence-level likelihoods for long LLM outputs can become vanishingly small (see Figure 1b).

Existing works address only parts of these challenges. Prediction-powered inference (PPI) (Angelopoulos et al., 2023a,b) provides a semi-supervised framework that augments limited ground-truth labels with proxy predictions for unlabeled data. Yet, PPI-type methods typically assume labeled and unlabeled responses follow the same distribution and thus do not handle *distribution shift*. Another line of work, importance-weighting-based methods, including importance sampling (IS) (Precup et al., 2000; Liu et al., 2018) and doubly robust (DR) estimators (Dudík et al., 2011; Jiang and Li, 2016; Thomas and Brunskill, 2016), can in principle correct for such shift using likelihood or importance ratios. In LLM settings, however, long responses make sequence-level ratios difficult to estimate and highly unstable. Under weak behavior–target overlap and limited samples, the correction terms can degenerate, leading to *substantial bias and high variance* (see Figure 1c).

To address these challenges simultaneously, we propose the **O**ptimal **T**ransport-based **R**obust **O**ff-**P**olicy **E**valuation (**OTROPE**), a likelihood-free framework for robust off-policy evaluation of LLMs under behavior–target distribution shift. The key insight is to perform distributional correction in a *semantic space* induced by embedding functions, rather than in the original token-level sequence space. In this space, the behavior and target distributions can exhibit substantially greater overlap due to semantic similarity, even when their overlap over response sequences is weak. We then leverage this structure by using optimal transport (OT) (Villani et al., 2009) to align labeled behavior-policy samples with unlabeled target-policy samples. The resulting OT weights reweight labeled residuals according to their semantic relevance to the target, and are combined with proxy predictors to produce a robust semi-supervised evaluator for LLMs.

Our **contributions** are four-fold:

• Conceptually, we formulate model-shifted evaluation on the performance of LLM of interest as a semi-supervised OPE problem that jointly addresses limited human annotation and behavior–target distribution shift in the complex LLM settings.
• Methodologically, we propose OTROPE that corrects behavior–target distribution shift in a semantic embedding space using OT-based reweighting. Unlike likelihood-ratio-based OPE methods, OTROPE does not require token-level likelihoods, behavior-policy modeling, or sequence-level density-ratio estimation, while enjoys a *doubly robust*-style correction for proxy-based evaluation.
• Theoretically, we derive the bias of the PPI estimator under distribution shift, and show the finite-sample degeneracy, bias, and variance inflation of the DR estimator in OPE for LLMs. We then establish the *consistency* and *convergence rate* of OTROPE when either the reweighted behavior distribution or the proxy predictor converges.
• Empirically, we evaluate OTROPE on synthetic data and real LLM evaluation tasks. OTROPE consistently outperforms baseline methods. In particular, it enables *ensembles of weaker LLM*

evaluators to approach and sometimes *outperform stronger LLM judges*, demonstrating the practical value of optimal transport-based correction for scalable LLM evaluation.

Related Work. We briefly discuss recent work on OPE for LLMs and highlight how our setting differs and our contributions. Bhargava et al. (2024) formulated OPE from logged human feedback, with an emphasis on evaluating preference rankings rather than the quality of response-level generation. More recent studies (Wu et al., 2024; Huang et al., 2025; Xu et al., 2026) use OPE-inspired estimators mainly for constructing training objectives for alignment rather than for model evaluation. In contrast, we study *response-level LLM evaluation under model shift with high-dimensional text responses*. Inspired by OT (Villani et al., 2009; Courty et al., 2016; Redko et al., 2017; Reygner and Touboul, 2022), we construct a reweighting scheme by minimizing the Wasserstein distance between the behavior distribution and the target distribution. This allows us to correct human-labeled residuals, yielding *a stable correction without requiring explicit density or likelihood ratios*. Unlike prior uses of OT for distributionally robust optimization over uncertainty sets (Shen et al., 2023) or preference alignment in LLMs (Melnyk et al., 2024), our method is the first work on likelihood-free OPE using OT in LLM settings. We provide a more comprehensive review on OPE, PPI, and OT in Appendix A.

2 Preliminaries

We first formulate the OPE problem in the context of evaluating LLMs, and then review related baselines, including the PPI and DR estimators, discussing their limitations in LLM settings.

Problem Setup. We denote the prompt space and the response space by $\mathcal{X}$ and $\mathcal{Y}$, respectively, and define the joint space as $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Suppose there exists a golden rater (e.g., a human annotator) that provides ground-truth evaluations for prompt–response pairs $z = (x, y) \in \mathcal{Z}$. Formally, let $g^* : \mathcal{Z} \to \mathcal{G}$ denote the ground-truth evaluation function, where $\mathcal{G} \subset \mathbb{R}$ is the evaluation space.

We consider a labeled dataset $\mathcal{D}_1 = \{(x_i, y_i, g_i^*)\}_{i=1}^n$ of size n, where $x_i \sim \Phi$ denotes a prompt drawn from the distribution Φ, $y_i \sim \pi_0(\cdot|x_i)$ is a response generated by a *behavior policy*, and $g_i^* = g^*(x_i, y_i)$ represents the corresponding ground-truth evaluation. Let P denote the joint distribution of $z = (x, y)$ induced by $x \sim \Phi$ and $y \sim \pi_0(\cdot|x)$. Due to the high cost and time requirements of obtaining reliable human annotations, the labeled sample size n is typically *limited*. In contrast, we have access to a larger unlabeled dataset $\mathcal{D}_2 = \{(\widetilde{x}_j, \widetilde{y}_j)\}_{j=1}^N$ of size N, where $\widetilde{x}_j \sim \Phi$ and $\widetilde{y}_j \sim \pi(\cdot|\widetilde{x}_j)$, with π denoting the *target language model or policy*. Let Q denote the corresponding joint distribution of $\widetilde{z} = (\widetilde{x}, \widetilde{y})$. Ground-truth evaluations are *unavailable* for samples in $\mathcal{D}_2$. Throughout this work, we consider autoregressive language models, in which the likelihood of a response is factorized over token-level probabilities. Since $\mathcal{D}_2$ lacks ground-truth human evaluations, we adopt the *LLM-as-a-judge* paradigm to construct *proxy evaluators* that evaluate each prompt–response pair in $\mathcal{D}_2$, denoted as $\widehat{g}(\widetilde{z}_j)$ (with more details provided in Section 3.2). We further define the residual function $r(\widetilde{z}) = g^*(\widetilde{z}) - \widehat{g}(\widetilde{z})$. Our estimand is *the target policy value*:

$$V^*(\pi) = \mathbb{E}_{\widetilde{z}\sim Q}\left[g^*(\widetilde{z})\right] = \mathbb{E}_{\widetilde{x}\sim\Phi}\,\mathbb{E}_{\widetilde{y}\sim\pi(\cdot|\widetilde{x})}\left[g^*(\widetilde{x}, \widetilde{y})\right]. \tag{1}$$

Our goal is to estimate $V^*(\pi)$ based on the datasets $\mathcal{D}_1$ and $\mathcal{D}_2$, together with the predictor $\widehat{g}$.

Prediction-Powered Inference (PPI) Estimator. After obtaining the proxy evaluations, a natural approach is to estimate $V^*(\pi)$ directly using these predictions. This leads to the direct method (DM) (Li et al., 2010), defined as $\widehat{V}_{\mathrm{DM}}(\pi) = 1/N \sum_{j=1}^N \widehat{g}(\widetilde{z}_j)$. However, such proxy evaluations may be unreliable. To mitigate prediction error, PPI (Angelopoulos et al., 2023a) combines a small amount of labeled data with a larger set of unlabeled data with arbitrary proxy predictors. Specifically, we consider the following PPI-based estimator for the policy value:

$$\widehat{V}_{\mathrm{PPI}}(\pi) = \frac{1}{N}\sum_{j=1}^{N} \widehat{g}(\widetilde{z}_j) + \frac{1}{n}\sum_{i=1}^{n} \left(g^*(z_i) - \widehat{g}(z_i)\right), \tag{2}$$

where $z_i = (x_i, y_i) \in \mathcal{D}_1$ and $\widetilde{z}_j = (\widetilde{x}_j, \widetilde{y}_j) \in \mathcal{D}_2$. The second term serves as a correction term that accounts for the prediction error of $\widehat{g}$ using labeled data. Nevertheless, PPI assumes that $\mathcal{D}_1$ and $\mathcal{D}_2$ are drawn from the same distribution. Under *distribution shift*, i.e., $\pi_0 \neq \pi$, $\widehat{V}_{\mathrm{PPI}}(\pi)$ is unbiased only if the prediction error is invariant across distributions; otherwise, it becomes biased and unreliable.

Doubly Robust (DR) Estimator. To address distribution shift in OPE, the DR method (Jiang and Li, 2016) is widely adopted that combines outcome modeling with behavior–target density ratio

modeling (Precup et al., 2000) (i.e., π/π_0). Following the PPI framework, a DR-based value estimator is

$$\widehat{V}_{\mathrm{DR}}(\pi) = \frac{1}{N}\sum_{j=1}^{N}\widehat{g}(\widetilde{x}_j,\widetilde{y}_j) + \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(y_i|x_i)}{\pi_0(y_i|x_i)}\left(g^*(x_i,y_i) - \widehat{g}(x_i,y_i)\right), \tag{3}$$

where the inner part is the IS estimator $\widehat{V}_{\mathrm{IS}}(\pi) = (1/n)\sum_{i=1}^{n}\{\pi(y_i|x_i)/\pi_0(y_i|x_i)\}g^*(x_i,y_i)$. In the context of LLMs, sequence probabilities are products of token-level probabilities, so small differences can accumulate into weak overlap between behavior and target policies, leading to degenerate density ratios and unreliable importance weighting. Moreover, in black-box API settings, these likelihoods are often unavailable, further complicating policy evaluation.

3 Optimal Transport-based Robust Off-Policy Evaluation

In this section, we first use a toy example to illustrate the limitations of the PPI and DR methods in Section 3.1, where PPI can be biased under distribution shift, whereas DR can degenerate under weak overlap in high-dimensional LLM sequence spaces. This motivates our robust likelihood-free approach to response-level shift-aware residual correction, as presented in Section 3.2.

3.1 Toy Example

We generate the prompt covariate x from a truncated normal distribution. The behavior and target distributions, P and Q, are also parameterized as truncated normal distributions, with parameters given as functions of x. We utilize parameter $\tau \geq 0$ to control the strength of the distributional shift between P and Q. We use a misspecified predictor $\widehat{g}$, which induces heterogeneous residuals under P and Q. We set $n = 1000$ and $N = 2000$, and repeat the experiment over 200 independent replications. The detailed setup is provided in Section 5.1. As shown in Figure 1a, the overlap between the behavior and target distributions becomes very weak when the shift parameter $\tau = 2$. As a result, the density ratio is close to zero with high probability (Figure 1b), making the DR correction term negligible under finite sample sizes. Figure 1c reports the mean absolute error (MAE) of different estimators. Density ratio-based estimators, including IS and DR estimators, perform poorly, highlighting the ineffectiveness of importance weighting in this weak-overlap regime. Moreover, residual heterogeneity across P and Q leads to poor performance of PPI-based methods. Detailed results of MAE under varying distribution shift are reported in Appendix E.1.

Motivated by these observations, we propose a new weighting scheme that combines the residual-correction structure of PPI and DR with distributional reweighting in the embedding space. Figure 1a shows that our proposed method OTROPE reweights the behavior distribution to closely match the target distribution; moreover, Figure 1c demonstrates that OTROPE consistently achieves the lowest MAE as the distribution shift increases with τ, highlighting its robustness to shift.

3.2 Proposed Method: OTROPE

To address the negligible correction induced by near-zero density ratios, we consider an alternative approach based on OT to reweight the labeled residuals. Intuitively, OT seeks the most cost-efficient way to transform one probability measure into another. To define the transport cost between two prompt-response pairs, we shift from exact token-level distribution to a *semantic geometric distribution* framework by considering a semantic embedding function $e : \mathcal{Z} \to \mathbb{R}^d$, and denote its image by $e(\mathcal{Z}) = \{e(z) : z \in \mathcal{Z}\} \subseteq \mathbb{R}^d$. We denote by $e_{\#}P$ and $e_{\#}Q$ the pushforward measures of P and Q under e, respectively. Let $c(u,v) = \|u - v\|$, be the Euclidean cost on $\mathbb{R}^d$. The OT cost between the pushforward measures $e_{\#}P$ and $e_{\#}Q$ is defined as

$$\mathcal{T}_c(e_{\#}P, e_{\#}Q) \coloneqq \inf_{\gamma\in\Gamma(e_{\#}P, e_{\#}Q)}\int_{e(\mathcal{Z})\times e(\mathcal{Z})} c(u,v)\,\mathrm{d}\gamma(u,v),$$

where $\Gamma(e_{\#}P, e_{\#}Q)$ denotes the set of all couplings between $e_{\#}P$ and $e_{\#}Q$, i.e., the set of probability measures on $e(\mathcal{Z}) \times e(\mathcal{Z})$ with marginals $e_{\#}P$ and $e_{\#}Q$. The corresponding Wasserstein distance (Vaserstein, 1969), also known as the Earth Mover's Distance (Rubner et al., 2000), is defined as $W(\mu,\nu) \coloneqq \mathcal{T}_c(\mu,\nu)$. The Wasserstein distance admits a natural interpretation as the minimal transportation cost required to move probability mass from $e_{\#}P$ to $e_{\#}Q$ under the cost function c. In

Algorithm 1 OTROPE: Optimal Transport-based Robust Off-Policy Evaluation

Input: labeled dataset $\mathcal{D}_1 = \{(z_i, g_i^*)\}_{i=1}^n$ from the unknown behavior policy π_0, unlabeled dataset $\mathcal{D}_2 = \{\widetilde{z}_j\}_{j=1}^N$ from the target policy π, embedding map $e(\cdot)$, transport cost $c(\cdot,\cdot)$, and a predictor $\widehat{g}$, which is either pre-specified or trained on an $\mathcal{D}_1$.
Obtain predictions $\widehat{g}(z_i)$ for $z_i \in \mathcal{D}_1$ and $\widehat{g}(\widetilde{z}_j)$ for $\widetilde{z}_j \in \mathcal{D}_2$.
Construct embeddings for both samples: $\{e(z_i)\}_{i=1}^n$ and $\{e(\widetilde{z}_j)\}_{j=1}^N$.
Form the pairwise cost $c(e(z_i), e(\widetilde{z}_j))$, for $i = 1, \ldots, n$, $j = 1, \ldots, N$.
Optimize the OT weighting problem in (5) to obtain the weights $\boldsymbol{w}^*$.
Compute the $\widehat{V}_{\mathrm{OTROPE}}(\pi)$ by (6).
return $\widehat{V}_{\mathrm{OTROPE}}(\pi)$.

LLM evaluation settings, the pushforward measures $e_{\#}P$ and $e_{\#}Q$ are unknown and must be approximated from samples. Given empirical samples in $\mathcal{D}_1$ and $\mathcal{D}_2$, we define $e_{\#}P_n = 1/n \sum_{i=1}^n \delta_{e(z_i)}$, and $e_{\#}Q_N = 1/N \sum_{j=1}^N \delta_{e(\widetilde{z}_j)}$, where $\delta_{e(z)}$ denotes the Dirac measure at $e(z)$. Instead of relying on importance weight in the response space to address distribution shift, we propose to reweight the labeled samples in $\mathcal{D}_1$ within a semantic space such that the resulting weighted distribution aligns with the empirical distribution of the target samples in $\mathcal{D}_2$ under the Wasserstein distance. The resulting weighted residuals over $e_{\#}P_n$ can then serve as an approximation to the mean residual under $e_{\#}Q_N$. We formulate the weight estimation problem as an optimal transport minimization problem over reweighted empirical measures. Let $\Delta_n = \{\boldsymbol{w} \in \mathbb{R}^n \mid \sum_{i=1}^n w_i = 1,\ w_i \geq 0\}$ denote the probability simplex. We aim to determine optimal weights $\boldsymbol{w}^* = (w_1^*, \ldots, w_n^*)$ such that the reweighted labeled distribution closely matches the unlabeled distribution in terms of the Wasserstein distance. Specifically, we solve

$$\boldsymbol{w}^* \in \operatorname*{arg\,min}_{\boldsymbol{w} \in \Delta_n} W\left(\sum_{i=1}^n w_i\, \delta_{e(z_i)},\ \frac{1}{N}\sum_{j=1}^N \delta_{e(\widetilde{z}_j)}\right) = \operatorname*{arg\,min}_{\boldsymbol{w} \in \Delta_n} \mathcal{T}_c(e_{\#}P_{n,\boldsymbol{w}}, e_{\#}Q_N), \tag{4}$$

where $e_{\#}P_{n,\boldsymbol{w}} \coloneqq \sum_{i=1}^n w_i \delta_{e(z_i)}$ is the reweighted labeled distribution. This formulation seeks weights that minimize the transportation cost required to align the labeled behavior distribution with the target distribution under the embedding-induced geometry. In the empirical setting, this problem reduces to a finite-dimensional optimal transport problem over samples, given by

$$\min_{\boldsymbol{w} \in \Delta_n,\ \Gamma \in \mathbb{R}_+^{n \times N}} \sum_{i=1}^n \sum_{j=1}^N \Gamma_{ij}\, c\big(e(z_i), e(\widetilde{z}_j)\big), \quad \text{s.t. } \sum_{j=1}^N \Gamma_{ij} = w_i,\ \sum_{i=1}^n \Gamma_{ij} = \frac{1}{N}, \tag{5}$$

where $\Gamma \in \mathbb{R}_+^{n \times N}$ denotes the transport plan and Γ_{ij} represents the mass transported from z_i to $\widetilde{z}_j$. This is a linear program over $(\boldsymbol{w}, \Gamma)$, and can be efficiently solved using standard OT solvers or via entropic regularization (Cuturi, 2013). Replacing density ratios in Eq. (3) with these geometry-aware weights, we propose the **O**ptimal **T**ransport-based **R**obust **O**ff-**P**olicy **E**valuation (**OTROPE**):

$$\widehat{V}_{\mathrm{OTROPE}}(\pi) = \frac{1}{N}\sum_{j=1}^N \widehat{g}(\widetilde{z}_j) + \sum_{i=1}^n w_i^* \left(g^*(z_i) - \widehat{g}(z_i)\right). \tag{6}$$

Intuitively, as characterized by Eq. (4), a labeled sample z_i receives a larger weight w_i^* when it is closer in the semantic embedding space to regions where the target policy π places higher probability mass. As a result, its residual $g^*(z_i) - \widehat{g}(z_i)$ contributes more strongly to correcting the proxy prediction under the target distribution. We summarize OTROPE in Algorithm 1 and detail OT weights training in Appendix C.1. In practice, we can employ M small LLMs as evaluators, denoted by $\{g_i\}_{i=1}^M$, and aggregate their evaluations for each pair in $\mathcal{D}_2$ to construct $\widehat{g}$. We consider two aggregation strategies. The first is *mixture-of-experts* (MoE), which forms a weighted combination of judgments (detailed in Appendix C.2). The second is *majority voting* for pair preference, which uses the majority label among evaluators as the proxy prediction. We show in Section 5 that using OTROPE, *ensembles of weaker LLMs* can approach and sometimes *outperform stronger LLM judges*.

4 Theoretical Results

In this section, we first theoretically analyze the limitations of PPI and DR estimators, and then establish the consistency and convergence rate of the proposed OTROPE.

4.1 Limitations of PPI and DR Estimators

The following proposition shows that the PPI estimator in Eq. (2) is generally biased when there exists discrepancy between the behavior distribution P and the target distribution Q.

Proposition 1 (Bias of the PPI Estimator under Distribution Shift)**.** *If* $\mathbb{E}_{z\sim P}[r(z)] \neq \mathbb{E}_{z\sim Q}[r(z)]$*, then*

$$\left|\mathbb{E}\big[\widehat{V}_{\mathrm{PPI}}(\pi)\big] - V^*(\pi)\right| = |\mathbb{E}_{z\sim P}[r(z)] - \mathbb{E}_{z\sim Q}[r(z)]| > 0.$$

This follows immediately by taking expectation: $\mathbb{E}\big[\widehat{V}_{\mathrm{PPI}}(\pi)\big] = \mathbb{E}_{z\sim Q}[\widehat{g}(z)] + \mathbb{E}_{z\sim P}\big[g^*(z) - \widehat{g}(z)\big] = \mathbb{E}_{z\sim Q}[\widehat{g}(z)] + \mathbb{E}_{z\sim P}\big[r(z)\big]$. This result reflects an intrinsic limitation due to the discrepancy between P and Q. We next examine the limitations of density-ratio based DR estimator in Eq. (3).

Proposition 2 (Finite-sample degeneracy, bias, and variance inflation of DR)**.** *Denote the density ratio by* $\rho(z_i) = \pi(y_i|x_i)/\pi_0(y_i|x_i)$*. Suppose* $0 < c_r \le r(z) \le C_r$ *for all* $z \sim P$*. If there exist sequences* $a_n \in (0,1)$*,* $b_n > 0$*, and* $d_n \in (0,1)$*, such that* $\mathbb{P}\left(1/n\sum_{i=1}^n \rho(z_i) \le a_n\right) \ge 1 - d_n$*, and* $\mathbb{P}(|\widehat{V}_{\mathrm{DM}}(\pi) - V^*(\pi)| \ge b_n) \ge 1 - d_n$*, then,* $\left|\widehat{V}_{\mathrm{DR}}(\pi) - V^*(\pi)\right| \ge b_n - C_r a_n$*, with probability at least* $1 - 2d_n$*. If additionally* $|\operatorname{Cov}(\widehat{V}_{\mathrm{DM}}, R_n)| \le \kappa \operatorname{Var}(R_n)$ *with* $\kappa < 1/2$*, then*

$$\operatorname{Var}\{\widehat{V}_{\mathrm{DR}}(\pi)\} \ge (1 - 2\kappa)\left[c_r^2 \frac{(1-a_n)^2}{d_n} - C_r^2\right].$$

Proposition 2 formalizes the finite-sample failure of DR under weak overlap. When empirical density ratios are close to zero with high probability, the residual correction becomes negligible, leading to biased estimates. This degeneracy implies occasional large weights, resulting in high variance, which is frequently encountered in LLMs applications. Proofs are provided in Appendix B.1.

4.2 Consistency and Convergence Rate of OTROPE

In contrast, we show that the proposed OTROPE estimator consistently recovers the true value $V^*(\pi)$ under suitable conditions. We begin by introducing Assumption 1.

Assumption 1 (Overlap)**.** $\operatorname{supp}(e_{\#}Q) \subseteq \operatorname{supp}(e_{\#}P)$*.* $\operatorname{supp}(e_{\#}P)$ *and* $\operatorname{supp}(e_{\#}Q)$ *are compact.*

Assumption 1 is standard in the OT literature (Gangbo and McCann, 1996; Fournier and Guillin, 2015; Manole and Niles-Weed, 2024). In our setting, it can be interpreted as requiring overlap after dimensionality reduction via the embedding map, rather than in the original high-dimensional sequence space as in DR methods. This assumption is reasonable in LLM settings, where both P and Q arise from policies defined over the same prompt distribution and share a common generation space. Moreover, embedding representations of LLM-generated responses tend to concentrate within a bounded region of $\mathbb{R}^d$, further supporting the validity of this assumption.

Lemma 4.1. *Define* $\mathcal{M}_n := \left\{\sum_{i=1}^n w_i \delta_{e(z_i)} : w_i \ge 0,\ \sum_{i=1}^n w_i = 1\right\}$*. Suppose Assumption 1 holds. Then, as* $n \to \infty$*,*

$$\inf_{\mu\in\mathcal{M}_n} W(\mu, e_{\#}Q) \xrightarrow{a.s.} 0.$$

Lemma 4.1 shows that, as the sample size of $\mathcal{D}_1$ increases, reweighted distributions can approximate the target distribution Q well in the Wasserstein distance. The proof is provided in Appendix B.2.

Assumption 2 (Lipschitz Continuity)**.** *There exists a function* $\widetilde{r} : \mathbb{R}^d \to \mathbb{R}$ *such that* $r(z) = \widetilde{r}(e(z))$ *for all* $z \in \mathcal{Z}$*. Moreover, there exists a constant* $L_r > 0$ *such that* $\widetilde{r}$ *is* L_r*-Lipschitz, i.e.,* $|\widetilde{r}(u) - \widetilde{r}(v)| \le L_r\|u - v\|$ *for all* $u, v \in e(\mathcal{Z})$*.*

Assumption 2 is a standard regularity condition (Manole and Niles-Weed, 2024; Hundrieser et al., 2024; Melnyk et al., 2024), requiring the residual function to vary smoothly in the semantic space. The Lipschitz continuity ensures that samples that are close in the embedding space, and hence semantically similar, exhibit similar prediction errors, thereby enabling effective bias correction.

Theorem 4.2 (Consistency of OTROPE)**.** *Suppose Assumptions 1 and 2 hold. Then, as $n \to \infty$ and $N \to \infty$,*

$$\widehat{V}_{\mathrm{OTROPE}}(\pi) \xrightarrow{p} V^*(\pi).$$

Since the residual function $\widetilde{r}(e(z))$ is L_r-Lipschitz continuous, Lemma 4.1 implies that expectations of $\widetilde{r}$ under $e_{\#}P_{n,\boldsymbol{w}^*}$ and $e_{\#}Q$ become arbitrarily close. Consequently, the bias introduced by evaluating the correction term under $P_{n,\boldsymbol{w}^*}$ vanishes asymptotically. Detailed proofs are in Appendix B.3. Unlike the PPI estimator in Eq. (2), which uses the average residual under P and remains biased even under Assumptions 1 and 2, OTROPE in Eq. (6) reweights $e_{\#}P_n$ to better align with $e_{\#}Q_N$. Under a lower mass condition, we further establish the convergence rate of OTROPE.

Assumption 3 (Lower Mass Condition)**.** *There exist constants $c > 0$, $0 < d_0 \leq d$, and $\eta_0 > 0$ such that for all $u \in \mathrm{supp}(e_{\#}Q)$ and all $0 < \eta \leq \eta_0$, $e_{\#}P(B(u,\eta)) \geq c\eta^{d_0}$, where $e_{\#}P(B(u,\eta))$ denotes the probability mass assigned by the pushforward distribution $e_{\#}P$ to the Euclidean ball centered at u with radius η in the embedding space.*

Assumption 3 is a standard lower mass condition that guarantees sufficient local mass around the support of the target distribution, and is commonly used to derive convergence rates in OT analyses (Weed and Bach, 2019; Reygner and Touboul, 2022; Hundrieser et al., 2024).

Theorem 4.3 (Convergence Rate of OTROPE)**.** *Suppose Assumptions 1, 2 and 3 hold. If $Var_{\widetilde{z}\sim Q}(\widehat{g}(\widetilde{z}))$ is bounded, then*

$$\widehat{V}_{\mathrm{OTROPE}}(\pi) - V^*(\pi) = O_p\Big(N^{-1/2} + n^{-1/d_0} + N^{-1/d}\Big).$$

d_0 can be interpreted as the intrinsic dimension (Weed and Bach, 2019) of the support of the distribution in the embedding space, reflecting how the probability mass of $e_{\#}P$ concentrates locally. Moreover, the convergence rate we establish is at most $O_p(N^{-1/d})$, implying that embeddings on lower-dimensional manifolds yield faster convergence. This highlights the importance of selecting embedding representations that capture essential semantic structure while remaining low-dimensional.

More importantly, OTROPE enjoys a form of *double robustness*, as shown in the following theorem.

Theorem 4.4 (Double Robustness of OTROPE)**.** *Suppose $\|\widehat{g}_n - g^*\|_\infty = \sup_{z\in\mathcal{Z}} |\widehat{g}_n(z) - g^*(z)| = O_p(\alpha_n)$ with $\alpha_n \to 0$, and $\widehat{g}_n$ has bounded conditional variance under the target distribution. Then, for any possibly data-dependent nonnegative weights $\widehat{w}_i$ satisfying $\widehat{w}_i \geq 0$ and $\sum_{i=1}^n \widehat{w}_i = 1$, we have*

$$\widehat{V}_{\mathrm{OTROPE}}(\pi) - V^*(\pi) = O_p\left(N^{-1/2} + \alpha_n\right).$$

The double robustness follows by combining Theorem 4.3 with Theorem 4.4. As discussed earlier, when the optimal transport weights $\boldsymbol{w}^*$ solve Eq. (4) and the residual induced by $\widehat{g}$ satisfies Assumption 2, OTROPE is consistent and achieves the convergence rate stated in Theorem 4.3. Conversely, Theorem 4.4 shows that if the proxy predictor $\widehat{g}_n$ is sufficiently accurate, then OTROPE remains consistent even when the transport weights are not accurate, as long as they are nonnegative and normalized. This is because the weighted correction term is bounded by $O_p(\alpha_n)$ under the normalization of the weights. In particular, when $\alpha_n = O(n^{-1/d_0})$, OTROPE achieves the rate $O_p\left(N^{-1/2} + n^{-1/d_0}\right)$, which avoids the additional $N^{-1/d}$ term appearing in Theorem 4.3. Proofs of Theorems 4.3 and 4.4 are provided in Appendix B.4.

5 Experiments

We evaluate OTROPE on both controlled experiments with various distribution shifts and real application on LLM evaluations, with comparison to PPI-based and density-ratio-based baselines. Beyond PPI, IS, and DR introduced in Section 2, we further include PPI++ (Angelopoulos et al., 2023b). Baselines are detailed in Appendix D.1, and hyperparameter choices are detailed in Appendix D.4.

5.1 Controlled Experiments with Various Distribution Shifts

Settings. We first generate a covariate $x \sim \mathrm{TruncNormal}(0, 1^2; [-3, 3])$. Given x, we generate a d-dimensional response $y \in \mathbb{R}^d$, where we set $d = 5$. Under the behavior distribution P, given x_i, the coordinates of y_i are conditionally independent and satisfy $y_{ik}|x_i \sim$

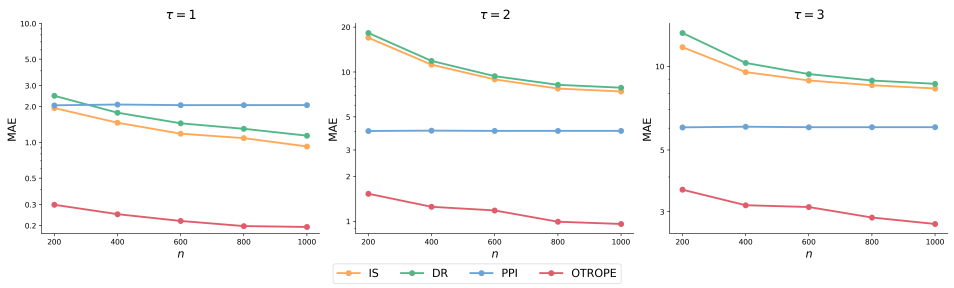


Figure 2: Log-scale mean absolute error under a fixed predictor $\widehat{g}$ for distribution shift levels $\tau \in \{1, 2, 3\}$. PPI++ is omitted since it nearly overlaps with PPI.

$\text{TruncNormal}(0.5\,x_i,\, 0.5^2; [-4, 4]), k = 1, \ldots, d$. Under the target distribution Q, we consider a distributional shift along a outcome-relevant direction $a \in \mathbb{R}^d$, where $\|a\| = 1$. Specifically, given $\widetilde{x}_j$, $\widetilde{y}_{jk}|\widetilde{x}_j \sim \text{TruncNormal}(0.5\,\widetilde{x}_j + \tau \times a_k,\, 0.5^2; [-4, 4]), k = 1, \ldots, d$. The parameter $\tau \geq 0$ controls the strength of the distributional shift between P and Q. The ground-truth outcome is given by $g^*(x, y) = 2 + 3x + 2a^\top y + \varepsilon, \varepsilon \sim N(0, 1^2)$. We consider a misspecified fixed predictor $\widehat{g}(x, y) = -x$ used in Section 3.1, which cannot capture the distributional shift induced by y, leading to biased estimates under the target distribution. Following Angelopoulos et al. (2023b), we fix $N = 2000$, vary the labeled sample size n, and repeat each experiment over 100 replications. We focus on distribution shift levels $\tau \in \{1, 2, 3\}$. We also consider a predictor $\widehat{g}(x, y)$ trained on $\mathcal{D}_1$, with details provided in Appendix E.2. For OTROPE, we use the embedding $e(x, y) = (x, a^\top y) \in \mathbb{R}^2$.

Results. Figure 2 reports the MAE for $\tau \in \{1, 2, 3\}$. In our experiments, PPI++ performs almost identically to PPI and is therefore omitted from the plots for clarity. Since the prediction model $\widehat{g}$ is fixed, the PPI-based methods exhibit little improvement as n increases. In contrast, OTROPE consistently outperforms all baselines across all settings. Under large distribution shifts, i.e., $\tau = 3$, estimation becomes more challenging, and more labeled samples are required for stable convergence. Nevertheless, OTROPE still achieves clear improvements over the baselines, demonstrating that the OT-based correction remains effective even in this difficult regime. Additional empirical checks are provided in the appendix: Appendix E.3 verifies Proposition 2, and Appendix E.4 reports an additional simulation under variance shift rather than mean shift.

5.2 Real Data Applications on LLM Evaluations

Datasets and models. REWARDBENCH 2 (Malik et al., 2026) is a multi-skill reward modeling benchmark containing model-generated responses and human preference annotations. CHATBOT ARENA (Zheng et al., 2023) provides large-scale human preference data from pairwise comparisons of LLM responses. We adapt both datasets to our evaluation setting as follows. We select two LLMs, or two pools of LLMs, to serve as the behavior and target policies, respectively, and construct an additional pool of LLMs as opponents. For each candidate response generated by either the behavior or target policy, we form a pairwise comparison against a response generated by an opponent model. Human preferences define the binary label g^*, where $g^* = 1$ indicates that the candidate response is preferred over the opponent response and $g^* = 0$ otherwise. The policy value is then defined as the expected preference win rate against the opponent pool. For each benchmark, we consider two target-policy settings. In REWARDBENCH 2, we consider a pooled target policy consisting of Claude 3.5 Sonnet and human responses, and a single-model target policy given by Claude 3.5 Sonnet. In CHATBOT ARENA, we use GPT-4 and GPT-3.5-Turbo as the two target policies. The behavior-policy configurations and more data construction details are provided in Appendix D.2. All experiments are conducted in a black-box model setting, where token-level probabilities are unavailable. Empirical support for Assumptions 1 and 3 over REWARDBENCH 2 and CHATBOT ARENA is provided in Appendices E.5 and E.6, respectively.

Implementation. We encode each prompt–response pair using bge-m3 embeddings (Chen et al., 2024b), and for pairwise comparisons, concatenate the embeddings of the candidate and opponent responses. Theorem 4.3 shows that the convergence rate depends on both the embedding dimension d and the intrinsic dimension d_0. This motivates the use of principal component analysis (PCA) (Hotelling, 1933; Abdi and Williams, 2010) before solving the OT problem: reducing the

Table 1: Mean absolute error of the estimated target policy value on REWARDBENCH 2 and CHATBOT ARENA. Standard deviations are reported in parentheses. The best result in each column is highlighted in **bold**. The Mixed policy is the mixture of Claude 3.5 Sonnet and human responses.

Method	**Benchmark**	REWARDBENCH 2		CHATBOT ARENA	
	Target Policy (π)	Mixed	Claude 3.5 Sonnet	GPT-4	GPT-3.5-Turbo
$\widehat{g}$ = *Mixture-of-Experts*	IS	0.263 (0.006)	0.273 (0.008)	0.414 (0.002)	0.272 (0.001)
	DR	0.129 (0.008)	0.125 (0.003)	0.088 (0.007)	0.065 (0.004)
	PPI	0.105 (0.006)	0.139 (0.001)	0.086 (0.007)	0.075 (0.005)
	PPI++	0.116 (0.003)	0.138 (0.001)	0.301 (0.004)	0.196 (0.003)
	OTROPE	**0.071** (0.046)	**0.103** (0.077)	**0.049** (0.018)	**0.010** (0.007)
$\widehat{g}$ = *Majority Voting*	IS	0.263 (0.006)	0.273 (0.008)	0.414 (0.002)	0.272 (0.001)
	DR	0.123 (0.006)	0.125 (0.007)	0.130 (0.001)	0.093 (0.002)
	PPI	0.097 (0.006)	0.122 (0.007)	0.124 (0.002)	0.108 (0.002)
	PPI++	0.118 (0.003)	0.135 (0.002)	0.365 (0.001)	0.234 (0.001)
	OTROPE	**0.066** (0.059)	**0.088** (0.083)	**0.029** (0.014)	**0.013** (0.012)
Stronger LLM-as-a-Judge	DeepSeek-V3.1	0.164	0.159	0.153	0.228

Table 2: Ablation study under the Mixed policy setting on REWARDBENCH 2, showing the effects of the correction term and proxy predictors across different labeled sample sizes n. DM with majority voting is invariant to n as as it does not use labeled samples; OT is identical across aggregation strategies as it does not use proxy predictions.

Estimator	Mixture-of-Experts			Majority Voting		
	$n = 800$	$n = 1200$	$n = 1600$	$n = 800$	$n = 1200$	$n = 1600$
DM	0.252 (0.009)	0.252 (0.007)	0.256 (0.006)	0.296 (0.000)	0.296 (0.000)	0.296 (0.000)
OT	0.272 (0.073)	0.236 (0.085)	0.258 (0.050)	0.272 (0.073)	0.236 (0.085)	0.258 (0.050)
OTROPE	**0.144** (0.048)	**0.114** (0.036)	**0.071** (0.046)	**0.099** (0.083)	**0.117** (0.067)	**0.066** (0.059)

embedding dimension can improve convergence and computational efficiency, while the low-intrinsic-dimensional structure suggests that the dominant components retain most of the relevant information. Since token probability of responses are not unavailable in the black-box setting, we use a classifier-based (CLS) importance weight estimator (Sugiyama et al., 2012) to approximate the density ratios in IS and DR (also see Appendix D.1). Each setting is repeated for 10 times.

Results. As shown in Table 1, OTROPE consistently achieves the lowest MAE across all datasets under different target policies and proxy evaluators. The larger standard deviations observed on REWARDBENCH 2 arise because the behavior and target policies are constructed as pools of LLMs rather than from a single LLM, resulting in hybrid behavior and target distributions. Baseline estimators may achieve lower variance but suffer from persistent large bias, whereas our method reduces this bias and yields more accurate estimates. The variance of OTROPE is also affected by the regularization term λ introduced later in Section 5.3, which controls a trade-off between matching stability and accuracy. Furthermore, OTROPE with small LLMs (all with <15B parameters, detailed in Appendix D.3) consistently outperforms all baselines and even surpasses the stronger standalone judge, DeepSeek-V3.1 (Liu et al., 2024) (with 671B parameters). This shows that *OT-based correction* can enable *ensembles of weaker evaluators* to match or *exceed much stronger models*.

5.3 Additional Results

Ablation study. We consider two ablations. First, we remove the correction term and use the DM estimator based only on proxy evaluations. Second, we remove the proxy evaluations and apply OT-based reweighting directly to the ground-truth evaluations, i.e., $\widehat{V}_{\mathrm{OT}}(\pi) = (1/n)\sum_{i=1}^{n} w_i^* \, g^*(x_i, y_i)$. We conduct the ablation study on the Mixed policy in REWARDBENCH 2 and summarize the results in Table 2. The relatively large errors of the DM estimator indicate that proxy evaluations alone can be unreliable. The superior performance of OTROPE highlights the importance of the correction term, which mitigates proxy misspecification and improves robustness. The OT estimator also performs worse than OTROPE because, when the proxy evaluator captures the dominant variation in the ground-truth evaluation, the residual $g^* - \widehat{g}$ is expected to vary more smoothly than g^* itself.

Sensitivity analysis of λ. In our method, λ is the entropic regularization parameter in the Sinkhorn iterations. It controls the trade-off between matching accuracy and computational stability; see Appendix C.1 for optimization details. Smaller λ gives more discriminative OT matching but is more computationally expensive, while larger λ improves efficiency but may reduce accuracy by smoothing pairwise cost differences. We conduct a sensitivity analysis on CHATBOT ARENA using GPT-4 as the target policy with MoE aggregation. The results in Table 3 show that stable performance across different sample sizes for $\lambda \in [0.05, 0.10]$. Smaller values of λ are preferred for more heterogeneous distributions, as they preserve finer pairwise cost structure and yield more accurate matching. For larger sample sizes, we increase λ within the empirically stable range to improve computational efficiency.

Table 3: Stable performance with different λ and sample size n.

λ	n		
	800	1200	1600
0.05	0.080 (0.025)	0.058 (0.028)	0.054 (0.015)
0.08	0.116 (0.073)	0.053 (0.061)	0.042 (0.024)
0.10	0.086 (0.060)	0.062 (0.025)	0.049 (0.018)

Sensitivity to the embedding representation. To assess whether OTROPE depends on a specific high-quality embedding model, we replace bge-m3 with a simple TF–IDF (Salton and Buckley, 1988) bag-of-words vectorizer using 1024 features and up to 2-grams, while keeping the rest of the pipeline unchanged. As shown in Table 4, OTROPE remains the best-performing estimator in three out of the four aggregation–target-policy combinations. These results suggest that OTROPE is reasonably robust to the choice of representation and remains effective even with a much simpler lexical embedding. Moreover, its performance is generally comparable to that obtained with bge-m3, as shown in Table 1.

Table 4: Mean absolute error of the estimated target policy value on CHATBOT ARENA using TF–IDF embeddings.

	Mixture-of-Experts		Majority Voting	
Estimator	GPT-4	GPT-3.5 Turbo	GPT-4	GPT-3.5 Turbo
IS	0.438 (0.004)	0.322 (0.003)	0.438 (0.003)	0.322 (0.003)
DR	0.023 (0.011)	**0.006 (0.005)**	**0.105 (0.004)**	0.036 (0.003)
PPI	0.028 (0.011)	0.023 (0.007)	0.124 (0.002)	0.108 (0.002)
PPI++	0.276 (0.004)	0.174 (0.003)	0.365 (0.001)	0.234 (0.001)
OTROPE	**0.013 (0.009)**	**0.006 (0.004)**	0.114 (0.009)	**0.007 (0.004)**

Table 5: Mean absolute error on CHATBOT ARENA under GPT-3.5-Turbo, with $n = 800$ labeled and $N = 1080$ unlabeled samples.

Estimator	Mixture-of-Experts	Majority Voting
IS	0.287 (0.011)	0.287 (0.011)
DR	0.077 (0.015)	0.091 (0.016)
PPI	0.092 (0.017)	0.113 (0.016)
PPI++	0.187 (0.016)	0.234 (0.012)
OTROPE	**0.045 (0.037)**	**0.083 (0.042)**

Evaluation in the label-scarce regime. To evaluate OTROPE in a real-data setting closer to the label-scarce regime considered in our simulations, we conduct the CHATBOT ARENA experiment with GPT-3.5-Turbo as the target policy, using $n = 800$ labeled behavior-policy samples and $N = 1080$ unlabeled target-policy samples. All other experimental configurations and hyperparameters follow those described above. The results are reported in Table 5. OTROPE achieves the lowest MAE under both aggregation schemes when $n < N$, indicating that the proposed OT-based correction remains effective in this label-scarce real-data setting.

Sensitivity to PCA dimension and transport cost. We further examine the robustness of OTROPE to two implementation choices: the PCA truncation dimension and the transport cost function. The corresponding sensitivity analyses are provided in Appendices E.7 and E.8, respectively.

6 Conclusion and Limitation

In this paper, we study LLM evaluation from an OPE perspective. We propose OTROPE, which uses optimal transport to reweight labeled residuals in a semantic embedding space. We prove consistency and convergence rates, and show that OTROPE consistently outperforms baselines on synthetic and real LLM evaluation tasks. Despite these advantages, OTROPE has several limitations. First, OT-based weighting can be computationally expensive; we use PCA for efficiency, but it may lose fine-grained information in dense embedding spaces. Second, our current formulation focuses on static input–output evaluation. Extending it to sequential evaluation settings, such as multi-turn interactions that can be modeled as bandits or Markov decision processes, is left for future work.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Science Foundation under Grant No. DMS-2401271 and by Thinking Machines Lab through the Tinker Research Grant. We thank the area chair and anonymous reviewers for their helpful comments.

References

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023a). Prediction-powered inference. *Science*, 382(6671):669–674.

Angelopoulos, A. N., Duchi, J. C., and Zrnic, T. (2023b). Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*.

Bhargava, A., Jain, L., Kveton, B., Liu, G., and Mukherjee, S. (2024). Off-policy evaluation from logged human feedback. *arXiv preprint arXiv:2406.10030*.

Boyeau, P., Angelopoulos, A. N., Yosef, N., Malik, J., and Jordan, M. I. (2024). Autoeval done right: Using synthetic data for model evaluation. *arXiv preprint arXiv:2403.07008*.

Cai, H., Lu, W., and Song, R. (2020). On validation and planning of an optimal decision rule with application in healthcare studies. In *International Conference on Machine Learning*, pages 1262–1270. PMLR.

Cai, H., Shi, C., Song, R., and Lu, W. (2021). Deep jump learning for off-policy evaluation in continuous treatment settings. *Advances in Neural Information Processing Systems*, 34:15285–15300.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620.

Chen, G. H., Chen, S., Liu, Z., Jiang, F., and Wang, B. (2024a). Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. (2024b). Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865.

Cowen-Breen, C., Agarwal, A., Bates, S., Cohen, W. W., Eisenstein, J., Globerson, A., and Fisch, A. (2026). Multiple-prediction-powered inference. *arXiv preprint arXiv:2603.27414*.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Datta, J. and Polson, N. G. (2025). Prediction-powered inference with inverse probability weighting. *arXiv preprint arXiv:2508.10149*.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104.

Fisch, A., Maynez, J., Hofer, R., Dhingra, B., Globerson, A., and Cohen, W. W. (2024). Stratified prediction-powered inference for effective hybrid evaluation of language models. *Advances in Neural Information Processing Systems*, 37:111489–111514.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738.

Fu, J., Ng, S. K., Jiang, Z., and Liu, P. (2024). Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.

Gangbo, W. and McCann, R. J. (1996). The geometry of optimal transportation.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Huang, C., Wu, J., Xie, Z., Xia, Y., Wang, R., Yu, T., Mitra, S., McAuley, J., and Yao, L. (2025). Pluralistic off-policy evaluation and alignment. *arXiv preprint arXiv:2509.19333*.

Hundrieser, S., Staudt, T., and Munk, A. (2024). Empirical optimal transport between different measures adapts to lower complexity. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 60, pages 824–846. Institut Henri Poincaré.

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pages 652–661. PMLR.

Kantorovich, L. and Rubinstein, G. S. (1958). On a space of totally additive functions. *Vestnik Leningrad. Univ*, 13:52–59.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31.

Malik, S., Pyatkin, V., Land, S., Morrison, J., Smith, N. A., Hajishirzi, H., and Lambert, N. (2026). Rewardbench 2: Advancing reward model evaluation. In *The Fourteenth International Conference on Learning Representations*.

Manole, T. and Niles-Weed, J. (2024). Sharp convergence rates for empirical optimal transport with smooth costs. *The Annals of Applied Probability*, 34(1B):1108–1135.

Melnyk, I., Mroueh, Y., Belgodere, B., Rigotti, M., Nitsure, A., Yurochkin, M., Greenewald, K., Navratil, J., and Ross, J. (2024). Distributional preference alignment of llms via optimal transport. *Advances in Neural Information Processing Systems*, 37:104412–104442.

Peyré, G. and Cuturi, M. (2019). *Computational optimal transport: With applications to data science*. Now Foundations and Trends.

Precup, D., Sutton, R. S., and Singh, S. (2000). Eligibility traces for off-policy policy evaluation.

Rachev, S. T. and Rüschendorf, L. (1998). *Mass Transportation Problems: Volume I: Theory*. Springer.

Rakotomamonjy, A., Flamary, R., Gasso, G., Alaya, M. E., Berar, M., and Courty, N. (2022). Optimal transport for conditional domain matching and label shift. *Machine Learning*, 111(5):1651–1670.

Redko, I., Habrard, A., and Sebban, M. (2017). Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 737–753. Springer.

Reygner, J. and Touboul, A. (2022). Reweighting samples under covariate shift using a wasserstein distance criterion. *Electronic Journal of Statistics*, 16(1):3278–3314.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121.

Saad-Falcon, J., Khattab, O., Potts, C., and Zaharia, M. (2024). Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Schroeder, K. and Wood-Doughty, Z. (2024). Can you trust llm judgments? reliability of llm-as-a-judge. *arXiv preprint arXiv:2412.12509*.

Shen, Y., Cai, H., and Song, R. (2024). Doubly robust interval estimation for optimal policy evaluation in online learning. *Journal of the American Statistical Association*, 119(548):2811–2821.

Shen, Y., Xu, P., and Zavlanos, M. M. (2023). Wasserstein distributionally robust policy evaluation and learning for contextual bandits. *arXiv preprint arXiv:2309.08748*.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.

Sutton, R. S., Barto, A. G., et al. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International conference on machine learning*, pages 2139–2148. PMLR.

Turrisi, R., Flamary, R., Rakotomamonjy, A., and Pontil, M. (2022). Multi-source domain adaptation via weighted joint distributions optimal transport. In *Uncertainty in artificial intelligence*, pages 1970–1980. PMLR.

Vaserstein, L. N. (1969). Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72.

Villani, C. et al. (2009). *Optimal transport: old and new*, volume 338. Springer.

Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648.

Wu, J., Li, X., Wang, R., Xia, Y., Xiong, Y., Wang, J., Yu, T., Chen, X., Kveton, B., Yao, L., et al. (2024). Ocean: Offline chain-of-thought evaluation and alignment in large language models. *arXiv preprint arXiv:2410.23703*.

Xu, E., Ye, K., Zhou, H., Zhu, L., Quinzan, F., and Shi, C. (2026). Doubly robust alignment for large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yang, L., Dan, I., Xu, Y., Shuohang, W., Xu, R., and Chenguang, Z. (2023). Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv: 2303.16634*.

Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., Gao, T., Geyer, W., Huang, C., Chen, P.-Y., et al. (2024). Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Appendix

The appendix contains additional related work, theoretical analysis, experimental details, and supplementary results. Appendix A provides a more comprehensive review of OPE, PPI, and OT. Appendix B presents theoretical results on the failure modes of PPI and DR, as well as the consistency, convergence rate, and double robustness of the proposed OTROPE estimator. Appendix C provides more details on the training procedure and implementation. Appendix D describes the data generation process, baseline estimators, LLMs used for evaluation, and hyperparameter choices. Finally, Appendix E reports additional experimental results.

A Related Work

Off-policy Evaluation. Off-policy evaluation (OPE) (Sutton et al., 1998) aims to estimate the value of a target policy prior to deployment, using historical data collected under a different behavior policy. This strategy is particularly important in domains where online evaluation is expensive, risky, or potentially harmful (Shen et al., 2024). Standard approaches to OPE include importance sampling (Precup et al., 2000; Liu et al., 2018), the direct method (DM) (Li et al., 2010), and doubly robust (DR) estimators (Dudík et al., 2011; Jiang and Li, 2016; Thomas and Brunskill, 2016; Cai et al., 2020, 2021). Several recent works have adapted OPE to LLMs, but they address problems different from ours. Bhargava et al. (2024) study OPE from logged human feedback, with a focus on evaluating preference-ranking quality rather than raw generated responses. Wu et al. (2024); Huang et al. (2025); Xu et al. (2026) use OPE-inspired estimators primarily for alignment training, where the goal is to construct optimization objectives rather than to estimate the evaluation score of a target model. In contrast, we focus on response-level LLM evaluation under model shift, where the evaluation objects are high-dimensional text responses whose distributions shift across models.

Prediction-Powered Inference. Prediction-Powered Inference (PPI) (Angelopoulos et al., 2023a) is a class of semi-supervised methods that uses black-box model predictions as proxy variables to improve the efficiency and validity of classical statistical inference, closely related to difference estimators (Cassel et al., 1976; Särndal et al., 2003). Several extensions have been proposed within this framework. PPI++ (Angelopoulos et al., 2023b), for example, introduces a weighting parameter on autorater predictions to improve efficiency, while StratPPI (Fisch et al., 2024) considers settings where autorater performance varies across subpopulations or conditional distributions of the target data. PPI-type methods have also been applied to LLM evaluation (Boyeau et al., 2024; Saad-Falcon et al., 2024; Fisch et al., 2024; Cowen-Breen et al., 2026). However, these approaches typically assume that labeled and unlabeled responses are generated by the same model, and thus do not address settings where new models are evaluated using labeled historical responses from earlier models. Some prior works, including the original PPI literature, has considered distribution shift (Angelopoulos et al., 2023a; Datta and Polson, 2025), but only utilize raw importance weight.

Optimal Transport. Optimal Transport (OT) (Rachev and Rüschendorf, 1998; Villani et al., 2009) provides a mathematical framework for comparing two probability distributions by seeking the most cost-efficient way to transport mass from one distribution to another, with transportation costs specified through an underlying ground metric. The resulting Wasserstein distance (Vaserstein, 1969), also referred to as the Earth Mover's Distance (Rubner et al., 2000), or the Kantorovich–Rubinstein distance (Kantorovich and Rubinstein, 1958), has played a central role in probability and statistics, and has increasingly become a powerful tool in modern machine learning and data science (Peyré and Cuturi, 2019). Recently, OT has been increasingly used to address distributional discrepancies in machine learning, including covariate shift (Reygner and Touboul, 2022) and domain adaptation (Courty et al., 2016; Redko et al., 2017; Rakotomamonjy et al., 2022; Turrisi et al., 2022). Although these works also leverage OT-based discrepancies or reweighting strategies, their objectives differ substantially from ours: they primarily focus on distribution alignment for supervised prediction, whereas our goal is OPE under distribution shift for LLMs.

B Theoretical Analysis

B.1 Proof of Proposition 2

Proof. Let $T_n = \frac{1}{n}\sum_{i=1}^n \rho(z_i)$ and $R_n = \frac{1}{n}\sum_{i=1}^n \rho(z_i)r(z_i)$. Then $\widehat{V}_{\mathrm{DR}}(\pi) = \widehat{V}_{\mathrm{DM}}(\pi) + R_n$.

Bias lower bound. Consider the event $\mathcal{E}_1 = \{T_n \le a_n\}$. By assumption, $\mathbb{P}(\mathcal{E}_1) \ge 1 - d_n$. On this event $\mathcal{E}_1$, using $\rho(z_i) \ge 0$ and $|r(z_i)| \le C_r$, we have

$$|R_n| = \left|\frac{1}{n}\sum_{i=1}^n \rho(z_i)r(z_i)\right| \le \frac{1}{n}\sum_{i=1}^n |\rho(z_i)r(z_i)| \le \frac{1}{n}\sum_{i=1}^n \rho(z_i)C_r = C_r T_n \le C_r a_n.$$

Therefore, with probability at least $1 - d_n$, $|R_n| \le C_r a_n$.

Denote the event $\mathcal{E}_2 = \left\{\left|\widehat{V}_{\mathrm{DM}}(\pi) - V^*(\pi)\right| \ge b_n\right\}$. By assumption, $\mathbb{P}(\mathcal{E}_2) \ge 1 - d_n$. There is

$$\left|\widehat{V}_{\mathrm{DR}}(\pi) - V^*(\pi)\right| = \left|\widehat{V}_{\mathrm{DM}}(\pi) + R_n - V^*(\pi)\right| \ge \left|\widehat{V}_{\mathrm{DM}}(\pi) - V^*(\pi)\right| - |R_n|,$$

where the last inequality follows from the reverse triangle inequality.

On the event $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\left|\widehat{V}_{\mathrm{DR}}(\pi) - V^*(\pi)\right| \ge b_n - C_r a_n.$$

Finally, by the union bound,

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \ge 1 - \mathbb{P}(\mathcal{E}_1^c) - \mathbb{P}(\mathcal{E}_2^c) \ge 1 - d_n - d_n = 1 - 2d_n.$$

Hence, with probability at least $1 - 2d_n$,

$$\left|\widehat{V}_{\mathrm{DR}}(\pi) - V^*(\pi)\right| \ge b_n - C_r a_n.$$

Variance lower bound. We first lower bound the variance of R_n. Since $C_r > r(z) \ge c_r > 0$,

$$R_n = \frac{1}{n}\sum_{i=1}^n \rho(z_i)r(z_i) \ge c_r T_n, \qquad R_n \le C_r T_n.$$

Therefore,

$$\mathbb{E}[R_n^2] \ge c_r^2 \mathbb{E}[T_n^2], \qquad (\mathbb{E}[R_n])^2 \le C_r^2 (\mathbb{E}[T_n])^2.$$

Since $\mathbb{E}[T_n] = 1$,

$$\mathrm{Var}(R_n) = \mathbb{E}[R_n^2] - (\mathbb{E}[R_n])^2 \ge c_r^2 \mathbb{E}[T_n^2] - C_r^2.$$

It remains to lower bound $\mathbb{E}[T_n^2]$. By assumption, $\mathbb{P}(\mathcal{E}_1^c) \le d_n$. We have

$$1 = \mathbb{E}[T_n] = \mathbb{E}[T_n \mathbf{1}_{\mathcal{E}_1}] + \mathbb{E}[T_n \mathbf{1}_{\mathcal{E}_1^c}] \le a_n + \mathbb{E}[T_n \mathbf{1}_{\mathcal{E}_1^c}],$$

which implies

$$\mathbb{E}[T_n \mathbf{1}_{\mathcal{E}_1^c}] \ge 1 - a_n.$$

By Cauchy–Schwarz inequality,

$$\mathbb{E}[T_n \mathbf{1}_{\mathcal{E}_1^c}] \le \{\mathbb{E}[T_n^2]\}^{1/2} \{\mathbb{P}(\mathcal{E}_1^c)\}^{1/2} \le \{\mathbb{E}[T_n^2]\}^{1/2} d_n^{1/2}.$$

Therefore,

$$\mathbb{E}[T_n^2] \ge \frac{(1 - a_n)^2}{d_n}.$$

Combining the above bounds gives

$$\mathrm{Var}(R_n) \ge c_r^2 \frac{(1 - a_n)^2}{d_n} - C_r^2.$$

Finally,

$$\mathrm{Var}\{\widehat{V}_{\mathrm{DR}}(\pi)\} = \mathrm{Var}(\widehat{V}_{\mathrm{DM}}(\pi)) + \mathrm{Var}(R_n) + 2\,\mathrm{Cov}(\widehat{V}_{\mathrm{DM}}(\pi), R_n).$$

Using

$$|\,\mathrm{Cov}(\widehat{V}_{\mathrm{DM}}, R_n)| \le \kappa\,\mathrm{Var}(R_n),$$

we obtain

$$\mathrm{Var}\{\widehat{V}_{\mathrm{DR}}(\pi)\} \ge (1 - 2\kappa)\,\mathrm{Var}(R_n).$$

Substituting the lower bound for $\mathrm{Var}(R_n)$ completes the proof. □

B.2 Proof of Lemma 4.1

The support of $e_{\#Q}$ is defined as: $e(z) \in \mathrm{supp}(e_{\#Q})$ if and only if $e_{\#}Q(B(e(z), \varepsilon)) > 0$ for all $\varepsilon > 0$, where $B(e(z), \varepsilon)$ denotes the open ball centered at $e(z)$ under the embedding space and where $e_{\#}Q(\cdot)$ denotes the probability mass in distribution $e_{\#}Q$.

Proof. Let $S \coloneqq \mathrm{supp}(e_{\#}Q)$, which is compact by Assumption 1. For any $\varepsilon > 0$.

Since S is compact, there exist finitely many points $u_1, \ldots, u_m \in S$ such that

$$S \subseteq \bigcup_{k=1}^{m} B(u_k, \varepsilon),$$

where $B(u, \varepsilon)$ denotes the ball centered at u with radius ε under the embedding-induced distance, i.e.,

$$B(u, \varepsilon) = \{u' : \|u' - u\| < \varepsilon\}.$$

Because $u_k \in \mathrm{supp}(e_{\#}Q) \subseteq \mathrm{supp}(e_{\#}P)$, each ball $B(u_k, \varepsilon)$ has strictly positive probability $e_{\#}P(B(u_k, \varepsilon)) > 0$, $k = 1, \ldots, m$. Define the event

$$A_{n,k} \coloneqq \bigcup_{i=1}^{n} \{e(z_i) \in B(u_k, \varepsilon)\}.$$

Then

$$\mathbb{P}(A_{n,k}^c) = \big(1 - P(B(u_k, \varepsilon))\big)^n.$$

Since $\mathbb{P}(B(u_k, \varepsilon)) > 0$, we have

$$\sum_{n=1}^{\infty} \mathbb{P}(A_{n,k}^c) < \infty.$$

for each k. Hence, by the Borel–Cantelli lemma,

$$\mathbb{P}(\limsup_{n \to \infty} A_{n,k}^c) = 0,$$

This implies that $A_{n,k}^c$ occurs only finitely often almost surely, i.e., there exists a finite N_k such that for all $n \geq N_k$, $A_{n,k}$ happens. Let

$$N_\varepsilon \coloneqq \max_{1 \leq k \leq m} N_k.$$

Then for every $n \geq N_\varepsilon$, each ball $B(u_k, \varepsilon)$ contains at least one sample point among $e(z_1), \ldots, e(z_n)$. For each $k = 1, \ldots, m$, choose one index $i_k \in \{1, \ldots, n\}$ such that

$$e(z_{i_k}) \in B(u_k, \varepsilon).$$

Next, construct a measurable partition $C_1, \ldots, C_m$ of S such that

$$C_k \subseteq B(u_k, \varepsilon), \qquad \bigcup_{k=1}^{m} C_k = S.$$

Define

$$\mu_{n,\varepsilon} \coloneqq \sum_{k=1}^{m} e_{\#}Q(C_k)\, \delta_{e(z_{i_k})},$$

where $e_{\#}Q(C_k)$ denotes the probability mass of the set C_k under the distribution $e_{\#}Q$. Since each $i_k \in \{1, \ldots, n\}$ and the coefficients are nonnegative and sum to one, we have $\mu_{n,\varepsilon} \in \mathcal{M}_n$. Consider the coupling that sends all mass in C_k to the point $e(z_{i_k})$. Then we have

$$W(\mu_{n,\varepsilon}, e_{\#}Q) \leq \sum_{k=1}^{m} \int_{C_k} \|u - e(z_{i_k})\| \, \mathrm{d}e_{\#}Q(u).$$

For any $u \in C_k$, since $C_k \subseteq B(u_k, \varepsilon)$ and $e(z_{i_k}) \in B(u_k, \varepsilon)$,

$$\|u - e(z_{i_k})\| \leq \|u - u_k\| + \|u_k - \boldsymbol{e}(z_{i_k})\| < 2\varepsilon.$$

Therefore,

$$W(\mu_{n,\varepsilon}, e_{\#}Q) \leq \sum_{k=1}^{m} \int_{C_k} 2\varepsilon \, \mathrm{d}Q(u) = 2\varepsilon \sum_{k=1}^{m} Q(C_k) = 2\varepsilon.$$

Since $\mu_{n,\varepsilon} \in \mathcal{M}_n$, it follows that

$$\inf_{\mu \in \mathcal{M}_n} W(\mu, e_{\#}Q) \leq 2\varepsilon$$

for all $n \geq N_\varepsilon$. Therefore, we derive

$$\inf_{\mu \in \mathcal{M}_n} W(\mu, e_{\#}Q) \xrightarrow{a.s.} 0.$$

□

B.3 Proof of Theorem 4.2

Proof. $r(z) = g^*(z) - \widehat{g}(z)$. We denote the reweighted distribution induced by OT (Eq. (4)) as

$$e_{\#}P_{n,\boldsymbol{w}^*} = \sum_{i=1}^{n} w_i \delta_{e(z_i)}.$$

We first show that $W(e_{\#}P_{n,\boldsymbol{w}^*}, e_{\#}Q) \to 0$. By the triangle inequality of the Wasserstein distance,

$$W(e_{\#}P_{n,\boldsymbol{w}^*}, e_{\#}Q) \leq W(e_{\#}P_{n,\boldsymbol{w}^*}, e_{\#}Q_N) + W(e_{\#}Q_N, e_{\#}Q). \tag{B.1}$$

Since $e_{\#}P_{n,\boldsymbol{w}^*}$ is defined in terms of the optimal solution

$$\boldsymbol{w}^* \in \arg\min_{\mu \in \mathcal{M}_n} W(\mu, e_{\#}Q_N),$$

we have

$$W(e_{\#}P_{n,\boldsymbol{w}^*}, e_{\#}Q_N) = \inf_{\mu \in \mathcal{M}_n} W(\mu, e_{\#}Q_N). \tag{B.2}$$

For any $\mu \in \mathcal{M}_n$, applying the triangle inequality again yields

$$W(\mu, e_{\#}Q_N) \leq W(\mu, e_{\#}Q) + W(e_{\#}Q, e_{\#}Q_N).$$

Taking the infimum over $\mu \in \mathcal{M}_n$, we obtain

$$\inf_{\mu \in \mathcal{M}_n} W(\mu, e_{\#}Q_N) \leq \inf_{\mu \in \mathcal{M}_n} W(\mu, e_{\#}Q) + W(e_{\#}Q, e_{\#}Q_N). \tag{B.3}$$

Combining (B.1), (B.2) and (B.3), we obtain

$$W(e_{\#}P_{n,\boldsymbol{w}^*}, e_{\#}Q) \leq \inf_{\mu \in \mathcal{M}_n} W(\mu, e_{\#}Q) + 2\, W(e_{\#}Q_N, e_{\#}Q).$$

Under Assumption 1 that $\mathrm{supp}(e_{\#}Q)$ is compact, which must be bounded, we invoke Theorem 1 of Fournier and Guillin (2015), for $p = 1$ and $d \geq 3$,

$$\mathbb{E}[W(e_{\#}Q_N, e_{\#}Q)] \leq CN^{-1/d},$$

where C is a constant related to p. We now apply Markov's inequality. For any $M > 0$,

$$\mathbb{P}\Big(W(e_{\#}Q_N, e_{\#}Q) > MN^{-1/d}\Big) \leq \frac{\mathbb{E}[W(e_{\#}Q_N, e_{\#}Q)]}{MN^{-1/d}} \leq \frac{CN^{-1/d}}{MN^{-1/d}} = \frac{C}{M}.$$

Hence, for any $\varepsilon > 0$, choosing $M = C/\varepsilon$ yields

$$\mathbb{P}\Big(W(e_{\#}Q_N, e_{\#}Q) > MN^{-1/d}\Big) \leq \varepsilon,$$

which is exactly the definition of

$$W(e_{\#}Q_N, e_{\#}Q) = O_p(N^{-1/d}).$$

Thus we have

$$W(e_{\#}Q_N, e_{\#}Q) \xrightarrow{p} 0.$$

By Lemma 4.1, $\inf_{\mu\in\mathcal{M}_n} W(\mu, e_{\#}Q) \xrightarrow{a.s.} 0$. Hence,

$$W(e_{\#}P_{n,\boldsymbol{w}^*}, e_{\#}Q) \xrightarrow{p} 0.$$

We now return to OTROPE:

$$\widehat{V}_{\mathrm{OTROPE}}(\pi) = \frac{1}{N}\sum_{j=1}^{N} \widehat{g}(\widetilde{z}_j) + \sum_{i=1}^{n} w_i^* r(z_i),$$

which can be rewritten as

$$\widehat{V}_{\mathrm{OTROPE}}(\pi) = \frac{1}{N}\sum_{j=1}^{N} \widehat{g}(z) + \int r(z)\,\mathrm{d}P_{n,\boldsymbol{w}^*}(z) = \frac{1}{N}\sum_{j=1}^{N} \widehat{g}(\widetilde{z}_j) + \int \widetilde{r}(e(z))\,\mathrm{d}e_{\#}P_{n,\boldsymbol{w}^*}(e(z)).$$

Since $W(e_{\#}P_{n,\boldsymbol{w}^*}, e_{\#}Q) \xrightarrow{p} 0$, and $\widetilde{r}$ is L_r-Lipschitz continuous in the embedding space under Assumption 2, the Kantorovich–Rubinstein duality implies that

$$\left|\int \widetilde{r}(e(z))\,\mathrm{d}e_{\#}P_{n,\boldsymbol{w}^*}(e(z)) - \int \widetilde{r}(e(z))\,\mathrm{d}e_{\#}Q(e(z))\right| \le L_r\, W(e_{\#}P_{n,\boldsymbol{w}^*}, e_{\#}Q) \xrightarrow{p} 0.$$

Moreover, for the first term of OTROPE, according to Weak Law of Large Numbers, we have

$$\frac{1}{N}\sum_{j=1}^{N} \widehat{g}(\widetilde{z}_j) \xrightarrow{p} \mathbb{E}_{z\sim Q}[\widehat{g}(z)].$$

Therefore,

$$\widehat{V}_{\mathrm{OTROPE}}(\pi) \xrightarrow{p} \mathbb{E}_{z\sim Q}[\widehat{g}(z)] + \mathbb{E}_{z\sim Q}[r(z)] = \mathbb{E}_{z\sim Q}[g^*(z)] = V^*.$$

□

B.4 Proofs of Theorems 4.3 and 4.4

Proof. As we just proofed in Appendix B.3,

$$W(e_{\#}P_{n,\boldsymbol{w}^*}, e_{\#}Q) \le \inf_{\mu\in\mathcal{M}_n} W(\mu, e_{\#}Q) + 2W(e_{\#}Q_N, e_{\#}Q).$$

We now control the first term with the additional assumption. For any $u \sim e_{\#}Q$, define its nearest neighbor among the source embeddings by

$$R_n(u) \coloneqq \min_{1\le i\le n} \|u - e(z_i)\|.$$

Using the nearest-neighbor projection of $e_{\#}Q$ onto $\{e(z_i)\}_{i=1}^n$, there exists some $\mu_n \in \mathcal{M}_n$ such that

$$W(\mu_n, e_{\#}Q) \le \mathbb{E}_{u\sim e_{\#}Q}[R_n(u)|z_1, \ldots, z_n].$$

This follows by constructing a transport plan that maps each point $u \sim e_{\#}Q$ to its nearest neighbor among $\{e(z_i)\}_{i=1}^n$, so that the transport cost is exactly the expected nearest-neighbor distance. Hence,

$$\inf_{\mu\in\mathcal{M}_n} W(\mu, e_{\#}Q) \le \mathbb{E}_{u\sim e_{\#}Q}[R_n(u)|z_1, \ldots, z_n].$$

Assumption 3 implies

$$\mathbb{P}(R_n(u) > \eta) = (1 - e_{\#}P(B(u,\eta)))^n \le (1 - c\eta^{d_0})^n \le \exp(-cn\eta^{d_0}) \tag{B.4}$$

for all sufficiently small r. Since $\mathrm{supp}(e_{\#}Q)$ is compact, it has finite diameter. Hence, there exists a constant $D < \infty$ such that $R_n(u) \le D$ almost surely. Using the tail-integral representation for nonnegative random variables, we have

$$\mathbb{E}[R_n(u)] = \int_0^\infty \mathbb{P}(R_n(u) > \eta)\ = \int_0^D \mathbb{P}(R_n(u) > \eta)\mathrm{d}\eta.$$

By (B.4),

$$\mathbb{E}[R_n(u)] \le \int_0^D \exp(-cn\eta^{d_0})\mathrm{d}\eta \le \int_0^\infty \exp(-cn\eta^{d_0})\mathrm{d}\eta \le \frac{1}{d_0}c^{-1/d_0}\Gamma\left(\frac{1}{d_0}\right)n^{-1/d_0}.$$

Therefore,

$$\mathbb{E}[R_n(u)] = O(n^{-1/d_0}).$$

By Markov's inequality,

$$\inf_{\mu\in\mathcal{M}_n} W(\mu, e_{\#}Q) = O_p(n^{-1/d_0}).$$

We also have shown $W(e_{\#}Q_N, e_{\#}Q) = O_p(N^{-1/d})$ in Appendix B.3. Combining the two bounds gives

$$W(e_{\#}P_{n,\boldsymbol{w}^*}, e_{\#}Q) = O_p(n^{-1/d_0} + N^{-1/d}).$$

We now decompose OTROPE:

$$\widehat{V}_{\mathrm{OTROPE}}(\pi) - V^*(\pi) = \left[\frac{1}{N}\sum_{j=1}^N \widehat{g}(\widetilde{z}_j) - \mathbb{E}_{z\sim Q}[\widehat{g}(z)]\right] + \left[\sum_{i=1}^n w_i^* r(z_i) - \mathbb{E}_{z\sim Q}[r(z)]\right]. \quad \text{(B.5)}$$

Since $\mathrm{Var}_{\widetilde{z}\sim Q}(\widehat{g}(\widetilde{z}))$ is bounded, by Chebyshev's inequality, the first term

$$\frac{1}{N}\sum_{j=1}^N \widehat{g}(\widetilde{z}_j) - \mathbb{E}_{z\sim Q}[\widehat{g}(z)] = O_p(N^{-1/2}).$$

For the second term,

$$\left|\int r(z)\,\mathrm{d}P_{n,\boldsymbol{w}^*}(z) - \int r(z)\,\mathrm{d}Q(z)\right| \le L_r W(e_{\#}P_{n,\boldsymbol{w}^*}, e_{\#}Q).$$

Thus,

$$\sum_{i=1}^n w_i^* r(z_i) - \mathbb{E}_{z\sim Q}[r(z)] = O_p(n^{-1/d_0} + N^{-1/d}).$$

Combining the two terms yields

$$\widehat{V}_{\mathrm{OTROPE}}(\pi) - V^*(\pi) = O_p\Big(N^{-1/2} + n^{-1/d_0} + N^{-1/d}\Big).$$

Additionally, suppose

$$\|\widehat{g}_n - g^*\|_\infty = O_p(\alpha_n), \qquad \alpha_n \to 0.$$

Then $\|r\|_\infty = O_p(\alpha_n)$. By the decomposition (B.5), the first term is $O_p(N^{-1/2})$ by Chebyshev's inequality under bounded variance. For the second term, since $w_i^* \ge 0$ and $\sum_i w_i^* = 1$,

$$\left|\sum_{i=1}^n w_i^* r(z_i)\right| \le \sum_{i=1}^n w_i^* |r(z_i)| \le \|r\|_\infty = O_p(\alpha_n),$$

and

$$|\mathbb{E}_{z\sim Q}[r(z)]| \le \mathbb{E}_{z\sim Q}[|r(z)|] \le \|r\|_\infty = O_p(\alpha_n).$$

Hence,

$$\sum_{i=1}^n w_i^* r(z_i) - \mathbb{E}_{z\sim Q}[r(z)] = O_p(\alpha_n).$$

Combining the two bounds gives

$$\widehat{V}_{\mathrm{OTROPE}}(\pi) - V^*(\pi) = O_p(N^{-1/2} + \alpha_n).$$

□

C Training and Implementation Details

C.1 Optimal Transport Weight Training

For ease of notation, let $E_1 = \{e_i\}_{i=1}^n = \{e(z_i)\}_{i=1}^n$, and $E_2 = \{\widetilde{e}_j\}_{j=1}^N = \{e(\widetilde{z}_j)\}_{j=1}^N$ denote the embedding representations of $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. If PCA is enabled, both E_1 and E_2 are transformed into a lower-dimensional OT space before computing the weights.

We estimate the OT weights by solving an entropy-regularized optimal transport problem (Cuturi, 2013) from the weighted empirical distribution on D_1 to the uniform empirical distribution on D_2 in (4). The ground cost is the normalized pairwise Euclidean distance between embeddings:

$$C_{ij} = \frac{\|e_i - \tilde{e}_j\|_2}{\max_{i,j} \|e_i - \tilde{e}_j\|_2}.$$

In implementation, the source weights are parameterized by unconstrained logits $\theta \in \mathbb{R}^n$:

$$w_i = \frac{\exp(\theta_i)}{\sum_{k=1}^n \exp(\theta_k)}.$$

This softmax parameterization ensures $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$ throughout optimization.

For a fixed $\boldsymbol{w}$, the entropy-regularized OT cost is computed using Sinkhorn iterations. Specifically, the target marginal is fixed as

$$b_j = \frac{1}{N}, \qquad j = 1, \ldots, N,$$

and the source marginal is $\boldsymbol{w}$. Define

$$M_{ij} = -\frac{C_{ij}}{\lambda},$$

where λ controls the strength of entropy regularization. Smaller values of λ lead to less regularized and more concentrated transport plans, whereas larger values yield smoother transport plans. The Sinkhorn algorithm alternates between updating two log-domain scaling variables $f \in \mathbb{R}^n$ and $g \in \mathbb{R}^N$:

$$f_i = \log w_i - \log \sum_{j=1}^N \exp(M_{ij} + h_j),$$

and

$$h_j = \log b_j - \log \sum_{i=1}^n \exp(M_{ij} + f_i).$$

These updates are repeated for a fixed number of inner Sinkhorn iterations, or until the change in f is below a numerical tolerance.

After the Sinkhorn iterations converge, the transport plan is recovered as

$$\Gamma_{ij} = \exp(f_i + h_j + M_{ij}).$$

The differentiable OT objective is then computed as

$$\langle \Gamma, C \rangle = \sum_{i=1}^n \sum_{j=1}^N \Gamma_{ij} C_{ij}.$$

The logits θ are initialized at zero, corresponding to uniform source weights. We then minimize the Sinkhorn OT cost using the Adam optimizer for a fixed number of outer iterations. In each outer iteration, the current logits determine $\boldsymbol{w}$, the inner Sinkhorn iterations compute the transport plan Γ, and the resulting OT cost is backpropagated to update θ. The final OT weights $\boldsymbol{w}^*$ are taken from the iterate with the smallest observed OT cost.

C.2 MoE Architecture and Optimization

We train the MoE model using a two-layer multilayer perceptron (MLP). For each input sample, the MLP takes the corresponding embedding vector as input and outputs a set of expert-combination weights. These weights are then used to aggregate the predictions from multiple fixed experts into a single MoE prediction.

The MoE model is trained on the labeled source data. The expert predictions are treated as fixed inputs, while the MLP learns the aggregation rule by matching the final MoE prediction to the observed label. For binary outcomes, we optimize the MLP parameters by minimizing the binary cross-entropy loss:

$$\mathcal{L}_{\mathrm{MoE}} = -\frac{1}{n}\sum_{i=1}^{n}\left[g^*(z_i)\log\widehat{g}(z_i) + \{1-g^*(z_i)\}\log\{1-\widehat{g}(z_i)\}\right],$$

where $\widehat{g}(z_i)$ denotes the MoE prediction for sample z_i, and $g^*(z_i)$ denotes the corresponding ground-truth binary label.

After training, we fix the MLP-based MoE model and use it to generate predictions on both the labeled source data and the target data. These predictions are saved and used in the downstream estimation step. Specifically, the target MoE predictions provide the plug-in estimate, while the source MoE predictions are used to construct the residual correction.

D Experimental Details

All experiments are conducted on a server equipped with an NVIDIA RTX A6000 GPU. Each repetition takes approximately 30 seconds to 5 minutes, depending on the sample size and target-policy setting.

D.1 Baseline Estimators

IS Estimator. The IS estimator (Precup et al., 2000) is defined as

$$\widehat{V}_{\mathrm{IS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(y_i|x_i)}{\pi_0(y_i|x_i)}\,g^*(x_i,y_i). \tag{D.1}$$

PPI++ Estimator. The PPI++ estimator (Angelopoulos et al., 2023b) is defined as

$$\widehat{V}_{\mathrm{PPI++}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}g^*(z_i) + \widehat{\lambda}\left(\frac{1}{N}\sum_{j=1}^{N}\widehat{g}(\widetilde{z}_j) - \frac{1}{n}\sum_{i=1}^{n}\widehat{g}(z_i)\right), \tag{D.2}$$

where

$$\widehat{\lambda} = \frac{\widehat{\mathrm{Cov}}_n\left(g^*(z_i),\widehat{g}(z_i)\right)}{(1+\frac{n}{N})\widehat{\mathrm{Var}}_{n+N}\left(\widehat{g}(z_i),\widehat{g}(\widetilde{z}_j)\right)}.$$

Here, $\widehat{\mathrm{Cov}}_n$ denotes the empirical covariance computed over the labeled samples $\{z_i\}_{i=1}^n$, and $\widehat{\mathrm{Var}}_{n+N}$ denotes the empirical variance of $\widehat{g}$ computed over the combined sample $\{z_i\}_{i=1}^n \cup \{\widetilde{z}_j\}_{j=1}^N$.

CLS Estimator for Density Ratio. When the densities $\pi(y|x)$ and $\pi_0(y|x)$ are unavailable, we estimate the importance weights using a probabilistic classifier. To this end, we construct a binary classification problem in which samples are drawn from a mixture of two conditional distributions: the target policy $\pi(\cdot|x)$ and the behavior policy $\pi_0(\cdot|x)$. Specifically, we assign label $l=1$ if $y\sim\pi(\cdot|x)$ and $l=0$ if $y\sim\pi_0(\cdot|x)$, with class priors $P(l=1)=\pi_1$ and $P(l=0)=\pi_0$. Following (Sugiyama et al., 2012), the density ratio can be expressed as

$$\frac{\pi(y|x)}{\pi_0(y|x)} = \frac{\pi_0}{\pi_1}\cdot\frac{P(l=1|y,x)}{P(l=0|y,x)}.$$

In our experiment, we apply

$$\frac{\pi_0}{\pi_1} = \frac{n}{N}.$$

We use this plug-in CLS density ratio estimator in IS and DR method in Section 5.2.

OT Estimator. We consider a direct OT-based estimator,

$$\widehat{V}_{\mathrm{OT}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} w_i^*\, g^*(x_i, y_i), \tag{D.3}$$

where w_i^* is the same as in OTROPE.

D.2 Data Construction

D.2.1 REWARDBENCH 2

Table D.1: Two policies evaluated in REWARDBENCH 2.

Policy Name	Claude 3.5 Sonnet	Mixed
Target Policy	`claude-3-5-sonnet-20241022`	`claude-3-5-sonnet-20241022`, `human`
Behavior Policy	`Llama-3.1-70B-Instruct`, `Qwen2.5-72B-Instruct`, `Qwen2.5-7B-Instruct`	`Llama-3.1-70B-Instruct`, `Qwen2.5-72B-Instruct`, `Qwen2.5-7B-Instruct`
Opponent Pool	`Llama-3.1-70B-Instruct`, `Llama-3.1-8B-Instruct`, `Llama-3.1-Tulu-3-8B`, `Mistral-7B-Instruct-v0.3`, `Qwen2.5-72B-Instruct`, `Qwen2.5-7B-Instruct`, `gpt-4o-2024-08-06`, `human`	`Llama-3.1-70B-Instruct`, `Llama-3.1-8B-Instruct`, `Llama-3.1-Tulu-3-8B`, `Mistral-7B-Instruct-v0.3`, `Qwen2.5-72B-Instruct`, `Qwen2.5-7B-Instruct`, `claude-3-5-sonnet-20241022`, `gpt-4o-2024-08-06`, `human`
Size $\mathcal{D}_1$	1,408	1,723
Size $\mathcal{D}_2$	493	591
Size (total)	1,901	2,314

We randomly sample n pairs from P to form $\mathcal{D}_1$, and use all pairs from Q as $\mathcal{D}_2$. We assume that the ground-truth preference g^* is observed for samples in $\mathcal{D}_1$ but unavailable for those in $\mathcal{D}_2$. The target value $V^*(\pi)$ is computed as the average ground-truth preference over $\mathcal{D}_2$. The true values are 0.843 for Claude 3.5 Sonnet and 0.870 for the Mixed policy. In each repetition, we set $n = 1400$ for Claude 3.5 Sonnet and $n = 1600$ for the Mixed policy in Section 5.2. Because the number of samples for each comparison pair in the original dataset is limited, we allow the target and behavior policies to also appear in the opponent pools, thereby constructing a larger opponent policy pool.

D.2.2 CHATBOT ARENA

Table D.2: Two policies evaluated in CHATBOT ARENA.

Policy Name	GPT-4	GPT-3.5-Turbo
Target Policy	`gpt-4`	`gpt-3.5-turbo`
Behavior Policy	`koala-13b`	`koala-13b`
Opponent Pool	`RWKV-4-Raven-14B`, `alpaca-13b`, `stablelm-tuned-alpha-7b`, `vicuna-13b`	`RWKV-4-Raven-14B`, `alpaca-13b`, `stablelm-tuned-alpha-7b`, `vicuna-13b`
Size $\mathcal{D}_1$	1,615	1,615
Size $\mathcal{D}_2$	998	1,080
Size (total)	2,613	2,695

The ground-truth values of the target policy are 0.755 for GPT-4 and 0.607 for GPT-3.5-Turbo. We randomly sample $n = 1600$ examples in Section 5.2.

D.3 LLM-as-a-Judge

We employ five LLMs to provide evaluation judgments: Phi-4-mini-instruct, gemma-7b-it, deepseek-llm-7b-chat, Llama-3.1-8B-Instruct, and Qwen2.5-14B-Instruct. These models are drawn from different families and may exhibit complementary strengths in evaluating diverse tasks. We also include a strong LLM, DeepSeek-V3.1, as a comparison.

D.4 Hyperpameters

Since REWARDBENCH 2 involves behavior and target policies induced by pools of LLMs, rather than by a single LLM as in CHATBOT ARENA, its induced data distribution is more complex. To

Table D.3: Hyperparameters for MoE training on REWARDBENCH 2 and CHATBOT ARENA.

Hyperparameter	REWARDBENCH 2	CHATBOT ARENA
Hidden dimensions	$(64, 32)$	$(64, 32)$
Optimizer	Adam	Adam
Learning rate	3×10^{-4}	1×10^{-3}
Batch size	256	256
Number of epochs	50	50
Temperature	$\{0.3, 0.5, 0.7, 1.0\}$	$\{0.3, 0.5, 0.7, 1.0\}$
Regularization	$\{0, 0.001, 0.01, 0.05, 0.1\}$	$\{0, 0.001, 0.01, 0.05, 0.1\}$
Training/validation split	$80\%/20\%$	$80\%/20\%$

improve optimization stability, we therefore use more training iterations and a smaller learning rate for REWARDBENCH 2.

We select the MoE hyperparameters via grid search over the regularization and temperature parameters (see Table D.3).

Table D.4: Hyperparameters used for OT weight optimization in the controlled experiments, under both fixed and trained $\widehat{g}$ settings.

Hyperparameter	Value / Description
Entropic regularization	0.1
Number of Sinkhorn iterations	1000
Outer optimizer	Adam
Number of outer iterations	500
Learning rate	0.05

Table D.5: Hyperparameters for OT weight optimization on REWARDBENCH 2 and CHATBOT ARENA.

Hyperparameter	Values
PCA dimension	128
Number of Sinkhorn iterations	2000
Outer optimizer	Adam
Number of outer iterations	1000
Learning rate	0.01

We set the entropic regularization parameter to $\lambda = 0.1$ for Claude 3.5 Sonnet and $\lambda = 0.08$ for the Mixed policy on REWARDBENCH 2. For CHATBOT ARENA, we use $\lambda = 0.05$ for GPT-4 and $\lambda = 0.1$ for GPT-3.5-Turbo. We choose λ within the empirically stable range $[0.05, 0.10]$ according to the matching difficulty of each target-policy setting and the computation efficiency. For more heterogeneous or distributionally distinct target policies, such as the Mixed policy and GPT-4, we use a smaller λ to preserve finer pairwise cost information and improve matching resolution. For smoother single-model settings, such as GPT-3.5-Turbo, we use the default value $\lambda = 0.10$, which provides better numerical smoothness and computational efficiency.

E Additional Experimental Results

E.1 Toy Example

We first provide additional numerical results in Table E.1 for the toy example in Section 3.1.

Table E.1: Mean absolute error (MAE) over 200 replications under varying distribution shift τ. Values are reported as mean with standard deviation in parentheses.

τ	IS	DR	PPI	PPI++	OTROPE
1	0.908 (1.077)	0.959 (1.186)	2.026 (0.126)	2.013 (0.063)	**0.236** (0.187)
2	6.906 (18.068)	6.420 (16.300)	4.005 (0.131)	3.998 (0.069)	**1.037** (0.703)
3	8.680 (8.832)	8.232 (5.112)	5.992 (0.138)	5.989 (0.070)	**2.880** (1.229)
4	9.865 (0.227)	9.878 (0.023)	7.886 (0.138)	7.894 (0.072)	**4.635** (1.740)
5	11.450 (0.000)	11.450 (0.022)	9.459 (0.118)	9.452 (0.073)	**6.205** (1.969)

E.2 More Results for Controlled Experiments

Here we show more results for the controlled experiments in Figure 2 of Section 5.1.

Table E.2: Mean absolute error and standard deviation of target policy value estimation with a fixed predictor. The labeled sample size n varies.

Estimator	n				
	200	400	600	800	1000
			$\tau = 1$		
IS	1.944 (5.290)	1.463 (2.548)	1.183 (1.724)	1.085 (1.395)	0.924 (1.208)
DR	2.464 (6.989)	1.776 (3.272)	1.446 (2.233)	1.300 (1.890)	1.139 (1.656)
PPI	2.048 (0.426)	2.079 (0.316)	2.055 (0.270)	2.057 (0.240)	2.057 (0.191)
PPI++	2.009 (0.120)	2.018 (0.087)	2.016 (0.076)	2.021 (0.068)	2.018 (0.061)
OTROPE	**0.299** (0.257)	**0.249** (0.181)	**0.219** (0.171)	**0.198** (0.164)	**0.195** (0.163)
			$\tau = 2$		
IS	16.966 (100.571)	11.194 (49.746)	8.919 (32.854)	7.744 (24.632)	7.403 (20.115)
DR	18.267 (112.970)	11.864 (55.908)	9.377 (36.970)	8.204 (27.769)	7.847 (22.789)
PPI	4.024 (0.288)	4.047 (0.210)	4.030 (0.178)	4.032 (0.157)	4.030 (0.122)
PPI++	3.999 (0.120)	4.007 (0.087)	4.006 (0.076)	4.010 (0.068)	4.008 (0.061)
OTROPE	**1.529** (0.711)	**1.250** (0.684)	**1.181** (0.673)	**0.993** (0.578)	**0.959** (0.613)
			$\tau = 3$		
IS	11.720 (37.407)	9.535 (17.959)	8.895 (11.463)	8.549 (8.226)	8.324 (6.286)
DR	13.186 (51.962)	10.286 (25.216)	9.378 (16.302)	8.891 (11.860)	8.652 (9.186)
PPI	6.033 (0.426)	6.064 (0.316)	6.040 (0.270)	6.042 (0.240)	6.042 (0.191)
PPI++	5.994 (0.120)	6.003 (0.087)	6.001 (0.076)	6.006 (0.068)	6.003 (0.061)
OTROPE	**3.601** (1.549)	**3.166** (1.489)	**3.119** (1.440)	**2.860** (1.373)	**2.711** (1.403)

We also report results for a predictor $\widehat{g}(x, y)$ trained on $\mathcal{D}_1$ in Table E.3 and Figure E.1. Specifically, $\widehat{g}(x, y)$ is implemented as a linear model trained on $\mathcal{D}_1$ using only the covariate x.

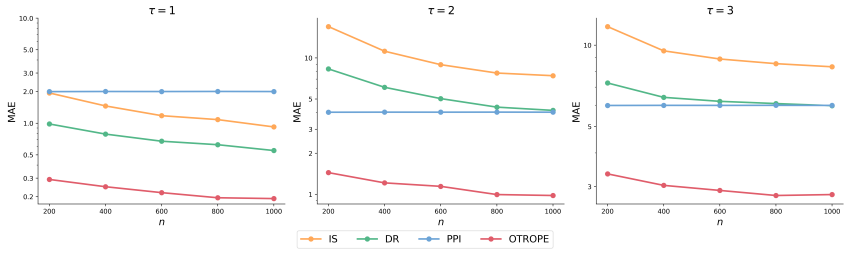


Figure E.1: Log-scale mean absolute error under the predictor $\widehat{g}$ trained on $\mathcal{D}_1$ for distribution shift levels $\tau \in \{1, 2, 3\}$. PPI++ is omitted since it nearly overlaps with PPI.

Together with Table E.2, these results show that OTROPE outperforms all baselines across the considered settings. This is consistent with Theorems 4.2 and 4.3, which impose no specific structural requirement on the relationship between the learned OT weights and the predictor $\widehat{g}$ or the induced residual function. Thus, $\widehat{g}$ can be an arbitrary model, including one trained on $\mathcal{D}_1$.

Table E.3: Mean absolute error and standard deviation of target policy value estimation with a predictor trained on $\mathcal{D}_1$. The labeled sample size n varies.

Estimator	n				
	200	400	600	800	1000
			$\tau = 1$		
IS	1.940 (5.291)	1.460 (2.550)	1.179 (1.726)	1.082 (1.398)	0.922 (1.210)
DR	0.984 (2.234)	0.788 (1.230)	0.674 (0.839)	0.624 (0.648)	0.549 (0.524)
PPI	1.998 (0.121)	2.003 (0.093)	2.001 (0.083)	2.005 (0.078)	2.001 (0.075)
PPI++	2.000 (0.120)	2.009 (0.087)	2.007 (0.076)	2.012 (0.068)	2.009 (0.061)
OTROPE	**0.291** (0.258)	**0.249** (0.177)	**0.218** (0.167)	**0.195** (0.160)	**0.192** (0.161)
			$\tau = 2$		
IS	16.967 (100.571)	11.196 (49.746)	8.921 (32.854)	7.745 (24.631)	7.405 (20.115)
DR	8.305 (36.887)	6.089 (18.829)	5.031 (12.417)	4.357 (9.182)	4.124 (7.273)
PPI	3.998 (0.121)	4.003 (0.093)	4.002 (0.083)	4.005 (0.078)	4.001 (0.075)
PPI++	4.000 (0.120)	4.009 (0.087)	4.008 (0.076)	4.012 (0.068)	4.010 (0.061)
OTROPE	**1.443** (0.701)	**1.214** (0.659)	**1.144** (0.657)	**0.996** (0.580)	**0.981** (0.573)
			$\tau = 3$		
IS	11.714 (37.408)	9.529 (17.960)	8.888 (11.465)	8.542 (8.228)	8.318 (6.287)
DR	7.247 (12.901)	6.413 (6.184)	6.199 (3.775)	6.083 (2.541)	5.973 (1.858)
PPI	5.985 (0.121)	5.991 (0.093)	5.989 (0.083)	5.992 (0.078)	5.988 (0.075)
PPI++	5.988 (0.120)	5.996 (0.087)	5.995 (0.076)	5.999 (0.068)	5.997 (0.061)
OTROPE	**3.344** (0.873)	**3.032** (0.885)	**2.904** (0.856)	**2.781** (0.763)	**2.804** (0.768)

E.3 Numerical Verification of Proposition 2

We provide a direct numerical verification of Proposition 2 using simulated data with oracle density ratios. We consider $n = 1000$ labeled samples and $N = 2000$ unlabeled samples, with 200 replications for each $\tau \in \{1, \ldots, 5\}$ and $d_n = 0.05$. We set

$$T_n = \frac{1}{n} \sum_{i=1}^{n} \rho(z_i)$$

and choose a_n as its empirical 95th percentile. We also set the direct-method bias $|\widehat{V}_{\mathrm{DM}} - V^*|$ to b_n at its empirical 5th percentile. These choices match the two high-probability conditions assumed in Proposition 2.

Bias bound. Among the replications where both conditions hold simultaneously, corresponding to the $1 - 2d_n$ guarantee, we compare the predicted lower bound $b_n - C_r a_n$ with the smallest observed value of $|\widehat{V}_{\mathrm{DR}} - V^*|$ among those replications. Table E.4 reports the results. The bound holds in all cases.

Table E.4: Numerical verification of the bias bound in Proposition 2. The constant $C_r = 16.51$ is a pooled upper bound on $|r(z)|$ across all τ.

τ	a_n	b_n	$b_n - C_r a_n$	$\min \lvert\widehat{V}_{\mathrm{DR}} - V^*\rvert$	Holds
1	1.373	3.965	-18.70	0.008	100%
2	2.719	5.959	-38.94	0.071	100%
3	0.177	7.948	5.03	6.119	100%
4	0.0006	9.847	9.838	9.847	100%
5	0.0000	11.415	11.415	11.415	100%

Variance bound. The variance lower bound in Proposition 2 additionally requires $0 < c_r \leq r(z)$. The original fixed predictor $\widehat{g}(x, y) = -x$ does not satisfy this condition. Therefore, for this diagnostic, we use a uniformly conservative predictor $\widehat{g}(x, y) = -x - 20$, which ensures $c_r > 0$, and compare the empirical variance $\mathrm{Var}(\widehat{V}_{\mathrm{DR}})$ with the predicted lower bound in Proposition 2. The results are shown in Table E.5. The bound is confirmed in the regimes $\tau = 1, 2, 3$.

The bound is cleanly confirmed in the regime targeted by the proposition, particularly for $\tau = 1, 2, 3$, where R_n still varies meaningfully across replications. For $\tau = 4, 5$, however, R_n is nearly identical across most of the 200 replications. As a result, the empirical quantiles and covariance terms used

Table E.5: Numerical verification of the variance lower bound in Proposition 2 under a conservative predictor satisfying $0 < c_r \leq r(z)$.

τ	$\widehat{\kappa}$	$\mathrm{Var}(\widehat{V}_{\mathrm{DR}})$	Predicted lower bound	Holds
1	0.000	30.5	-1.0×10^3	Yes
2	0.000	1.62×10^4	5.57×10^3	Yes
3	0.000	7.97×10^3	250	Yes

to estimate a_n, b_n, and $\widehat{\kappa}$ become unstable and no longer provide reliable empirical proxies for the population-level quantities appearing in Proposition 2.

E.4 Robustness under Variance Shift

Our main simulation varies the discrepancy between the behavior and target distributions primarily through a mean or location shift. This design is natural because the policy value is an expectation, and a location shift along an outcome-relevant direction provides a direct and interpretable way to control changes that affect the target policy value. However, this is only a choice of simulation design rather than a restriction of OTROPE. The OT objective aligns the full empirical distributions and does not assume that the behavior–target discrepancy arises only from the mean. Therefore, variance, scale, and more general distributional-shape shifts are also covered by the proposed framework.

To evaluate OTROPE under a variance or scale shift, we consider an additional simulation with $n = N = 1000$ and $d = 5$, where only the conditional scale changes while the conditional mean remains unchanged:

$$P:\ y_k \mid x \sim \mathrm{TruncNormal}(\gamma_k x, \sigma_{\mathrm{ref}}, [-4, 4]),\ Q:\ y_k \mid x \sim \mathrm{TruncNormal}(\gamma_k x, \sigma_{\mathrm{tgt}}, [-4, 4]).$$

We fix $\sigma_{\mathrm{ref}} = 0.5$ and consider two scale ratios, $\sigma_{\mathrm{tgt}}/\sigma_{\mathrm{ref}} = 0.5$ and 2.0. All other settings follow the toy example in the main paper. We use 20 replications and report the mean absolute error of the estimated target policy value in Table E.6; standard deviations are shown in parentheses.

Table E.6: Mean absolute error under variance-shift simulations. Standard deviations are reported in parentheses.

$\sigma_{\mathrm{tgt}}/\sigma_{\mathrm{ref}}$	IS	DR	PPI	OTROPE
0.50	0.226 (0.066)	0.287 (0.101)	0.083 (0.005)	0.104 (0.007)
2.00	2.163 (30.56)	2.138 (24.19)	0.120 (0.007)	0.125 (0.013)

OTROPE remains substantially more accurate and stable than IS and DR under both scale-shift settings. PPI performs slightly better in this particular experiment because the conditional mean structure is unchanged, making the setting especially favorable to its unweighted residual correction. When $\sigma_{\mathrm{tgt}}/\sigma_{\mathrm{ref}} = 2$, IS and DR become unstable because the target distribution has heavier tails than the behavior distribution; consequently, a small number of tail observations can receive very large importance weights and dominate the estimate. In contrast, PPI uses uniform weights, while OTROPE uses simplex-constrained OT weights rather than raw density ratios. Overall, the results indicate that OTROPE remains competitive and stable, and that its effectiveness is not limited to mean-shift settings.

E.5 Empirical Support for Assumption 1

Assumption 1 is a population-level support condition and cannot be formally verified from finite samples. Nevertheless, its plausibility can be empirically assessed through geometric coverage diagnostics in the embedding space. For each target-policy sample, we compare its nearest-neighbor distance to the behavior-policy samples with its nearest-neighbor distance within the target-policy samples. We then use the ratio between these two distances as a relative measure of support coverage. A ratio close to one suggests that the behavior-policy samples cover the corresponding target region at a similar local scale, whereas a larger ratio indicates weaker local support coverage.

Table E.7: Nearest-neighbor support coverage diagnostics in the embedding space. The ratio compares the target-to-behavior nearest-neighbor distance with the target-to-target nearest-neighbor distance. A ratio close to one indicates stronger local support coverage.

Dataset	Target → Behavior NN	Target → Target NN	Ratio	Fraction with ratio > 2
CHATBOT ARENA (GPT-4)	0.855	0.871	0.98	0.0%
CHATBOT ARENA (GPT-3.5-Turbo)	0.865	0.883	0.98	0.0%
REWARDBENCH 2 (Claude 3.5 Sonnet)	0.955	0.163	5.85	94.2%
REWARDBENCH 2 (Mixed)	0.936	0.164	5.71	94.1%

As shown in Table E.7, the ratios on CHATBOT ARENA are close to one, and no target samples have a ratio greater than two. This provides empirical support for Assumption 1 on CHATBOT ARENA. In contrast, REWARDBENCH 2 exhibits weaker overlap, with ratios around 5.7–5.9 and more than 94% of target samples having ratio greater than two. This indicates that the target samples are farther from the behavior-policy samples in the embedding space, which is consistent with the larger estimation errors observed on REWARDBENCH 2. Nevertheless, OTROPE still performs strongly relative to the competing baselines in this more challenging overlap regime.

E.6 Empirical Verification of Assumption 3

We empirically examine Assumption 3 using real data. Recall that Assumption 3 requires the local ball mass $e_{\#}P(B(u,\eta))$ to shrink no faster than η^{d_0}, where d_0 is the intrinsic dimension of the embedding support. To estimate d_0, for each target-policy sample we compute the fraction of behavior-policy samples that fall within radius η, average this quantity over all target-policy samples to obtain $\mathrm{mass}(\eta)$, and fit a linear regression between $\log \eta$ and $\log \mathrm{mass}(\eta)$. The fitted slope gives an estimate $\widehat{d}_0$, which measures the effective dimension of the local embedding geometry.

Table E.8: Estimated intrinsic dimension of the raw pre-PCA bge-m3 embeddings. The slope of the log–log regression gives $\widehat{d}_0$.

Dataset	$\widehat{d}_0$	R^2
CHATBOT ARENA (GPT-4)	23.98	1.0000
CHATBOT ARENA (GPT-3.5-Turbo)	24.01	1.0000
REWARDBENCH 2 (Mixed)	29.23	0.9996
REWARDBENCH 2 (Claude 3.5 Sonnet)	29.17	0.9997

Table E.8 reports the results on the raw pre-PCA bge-m3 embeddings. The nearly perfect log–log linearity, with $R^2 \geq 0.9996$ across all datasets, indicates that the power-law scaling in Assumption 3 closely matches the empirical embedding geometry. Moreover, the estimated intrinsic dimensions, $\widehat{d}_0 \approx 24$–29, are far below the raw embedding dimension of the concatenated bge-m3 representations and also below the PCA dimension used in our experiments. These findings provide empirical support for Assumption 3 and further justify the use of PCA for dimension reduction.

E.7 Sensitivity to the PCA Dimension

Here, we conduct a sensitivity analysis with respect to the embedding dimension. The bge-m3 embedding model produces fixed 1024-dimensional dense embeddings. For pairwise comparisons, we concatenate the embeddings of the candidate and opponent responses, resulting in 2048-dimensional representations. Since bge-m3 does not natively support adjustable embedding dimensions, we vary the effective embedding dimension through PCA truncation. In our main experiments, we use 128-dimensional PCA representations.

To assess the sensitivity of OTROPE to this dimension-reduction choice, we repeat the CHATBOT ARENA experiments using bge-m3 embeddings while reducing the PCA dimension from 128 to 64. All other hyperparameters and experimental configurations are kept unchanged. Table E.9 reports the results.

As expected, reducing the PCA dimension from 128 to 64 increases the MAE of OTROPE relative to the main results in Table 1, since a lower-dimensional representation retains less information and

Table E.9: Mean absolute error of the estimated target policy value on CHATBOT ARENA using 64-dimensional PCA-truncated bge-m3 embeddings. Standard errors are reported in parentheses.

	Mixture-of-Experts		Majority Voting	
Method	GPT-4	GPT-3.5	GPT-4	GPT-3.5
IS	0.408 (0.002)	0.261 (0.002)	0.408 (0.002)	0.261 (0.002)
DR	0.088 (0.009)	0.081 (0.007)	0.127 (0.002)	0.110 (0.002)
PPI	0.086 (0.009)	0.080 (0.007)	0.124 (0.002)	0.108 (0.002)
PPI++	0.302 (0.006)	0.197 (0.004)	0.365 (0.001)	0.234 (0.001)
OTROPE	**0.049 (0.013)**	**0.076 (0.007)**	**0.088 (0.012)**	**0.106 (0.006)**

yields a coarser pairwise-cost geometry for OT weighting. Nevertheless, OTROPE still achieves the lowest MAE across all four aggregation–target-policy combinations, indicating that the method remains effective under more aggressive PCA truncation.

E.8 Sensitivity to the Transport Cost

For the transport cost, we use the ℓ_2 distance in all main experiments. To assess sensitivity to this choice, we additionally evaluate OTROPE using the ℓ_1 distance on CHATBOT ARENA with GPT-3.5-Turbo as the target policy, while keeping all other hyperparameters and experimental configurations unchanged. Table E.10 reports the results.

Table E.10: Mean absolute error on CHATBOT ARENA under GPT-3.5-Turbo using the ℓ_1 transport cost. Standard deviations are reported in parentheses.

Method	Mixture-of-Experts	Majority Voting
IS	0.272 (0.001)	0.272 (0.001)
DR	0.069 (0.008)	0.094 (0.002)
PPI	0.079 (0.008)	0.108 (0.002)
PPI++	0.197 (0.004)	0.234 (0.001)
OTROPE	**0.025 (0.011)**	**0.034 (0.016)**

OTROPE achieves the lowest MAE under both aggregation schemes, indicating that its empirical performance is robust to the choice of transport cost. Its performance under the ℓ_1 cost is also close to that obtained with the default ℓ_2 cost, suggesting that the proposed correction is not sensitive to this specific metric choice.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We describe our method in the abstract and conclude the contributions in the introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations at the end of this paper.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: We state the assumptions for the theoretical results and provide complete proofs in Appendix B.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: In experiments section and Appendix D, we disclose the information of our experiments including both controlled experiments and real data applications on LLM evaluations.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We will release the code to reproduce our main experimental results under open source upon acceptance. During the review process the code is shared in supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: All the experimental details are provided in Appendix D.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We provide mean absolute error with the standard deviation under multiple replications.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Yes. We report the computational resources used for the experiments, including the type of GPU and its memory capacity in Appendix D.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics. It uses only publicly available datasets with proper citations, does not involve sensitive personal data, and releases code responsibly.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: The work is primarily methodological, focusing on the evaluation of LLMs. We do not identify direct positive or negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: This work does not release new datasets or models with high risk for misuse. It uses only publicly available datasets and focuses on LLM evaluation. Therefore, safeguards for high-risk releases are not applicable.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All datasets and models used in this work are publicly available and properly credited with appropriate citations. We adhere to the respective licenses and terms of use of these resources and ensure that they are used in accordance with their intended purposes.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

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

13. **New assets**

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

    Answer: [N/A]

    Justification: The work does not release any new datasets, and the proposed policy evaluation framework is not provided as a standalone asset.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A]

    Justification: This work does not involve any new crowdsourcing experiments or direct interaction with human subjects. All human preference data are obtained from existing publicly available datasets, and no new data collection, participant instructions, or compensation are involved.

    Guidelines:

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

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

    Answer: [N/A]

Justification: This work does not involve new studies with human participants. It relies solely on publicly available datasets with pre-collected human annotations, and no direct interaction with human subjects is conducted. Therefore, considerations such as participant risk, disclosure, or IRB approval are not applicable.

Guidelines:

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

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used only for writing, editing, or formatting purposes and do not affect the core methodology, scientific rigor, or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.