Doubly Robust Alignment for Large Language Models

Erhan Xu†[1], Kai Ye†[1], Hongyi Zhou†[2], Luhan Zhu[3], Francesco Quinzan‡[4], Chengchun Shi‡[1]
[1]London School of Economics, [2]Tsinghua University, [3]University of the Arts London, [4]University of Oxford
†equal contribution, ‡joint senior contributors

Problem: Model Misspecification in RLHF

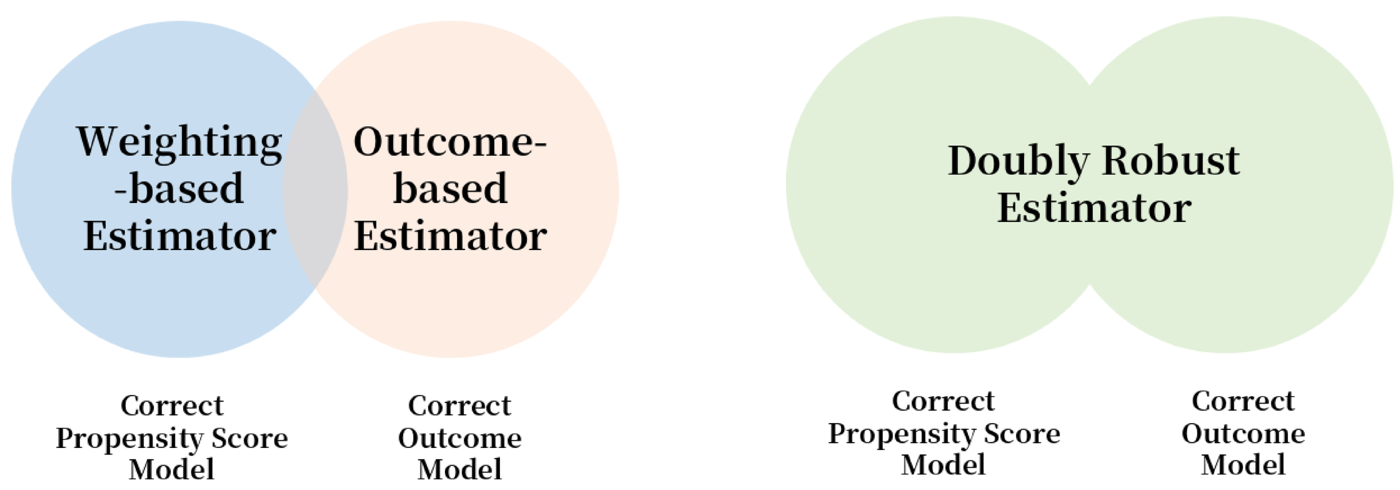
- **PPO-based algorithms:** Sensitive to reward model misspecification
  - Can lead to reward hacking and misguided policy learning
- **DPO-based algorithms:** Sensitive to reference policy misspecification
  - Performance degrades when reference policy is inaccurate
- **Preference-based algorithms:** Rely on correct preference model specification
  - Bradley-Terry (BT) model often violated due to intransitivity in human preferences

		Robust to misspecified:	preference model	reward model	reference policy
RLHF	Reward-based	PPO-type	✗	✗	✓
		DPO-type	✗	✓	✗
	Preference-based	IPO	✓	-	✗
		GPM	✗	-	✓
		DRPO	✓	✓	✓

Our Solution: Use Double Robust Method

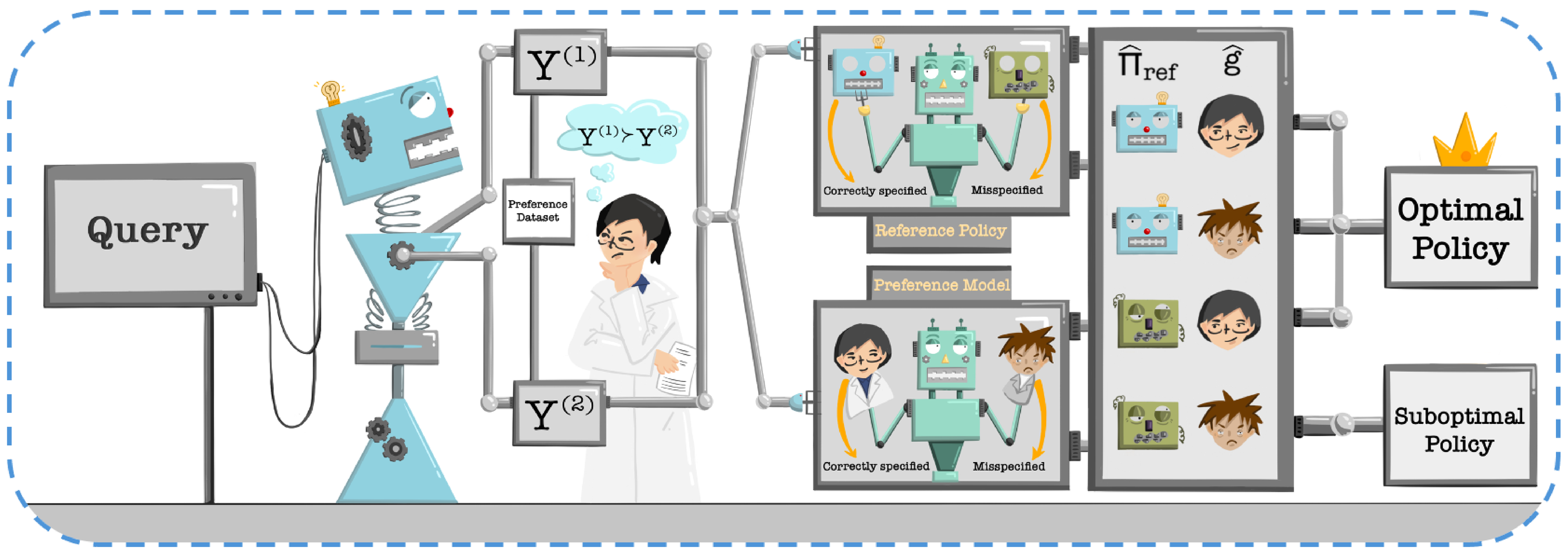
Doubly robust methods originate from the **missing data** and **causal inference** literature. Consider the estimation of **average treatment effect** (ATE) in causal inference. These methods estimate two models:

- A **propensity score** model for treatment assignment mechanism
- Similar to **reference policy** in LLMs
- An **outcome regression** model for subject's outcome given treatment
- Similar to **reward model** in LLMs



(a) Consistent only in each circle (b) Consistent in the union of circles

When DR methods meet LLMs:



Key Insight:

- Estimate two models: preference model $\widehat{g}$ and reference policy $\widehat{\pi}_{\text{ref}}$
- Construct estimator that remains consistent when *either* model is correct

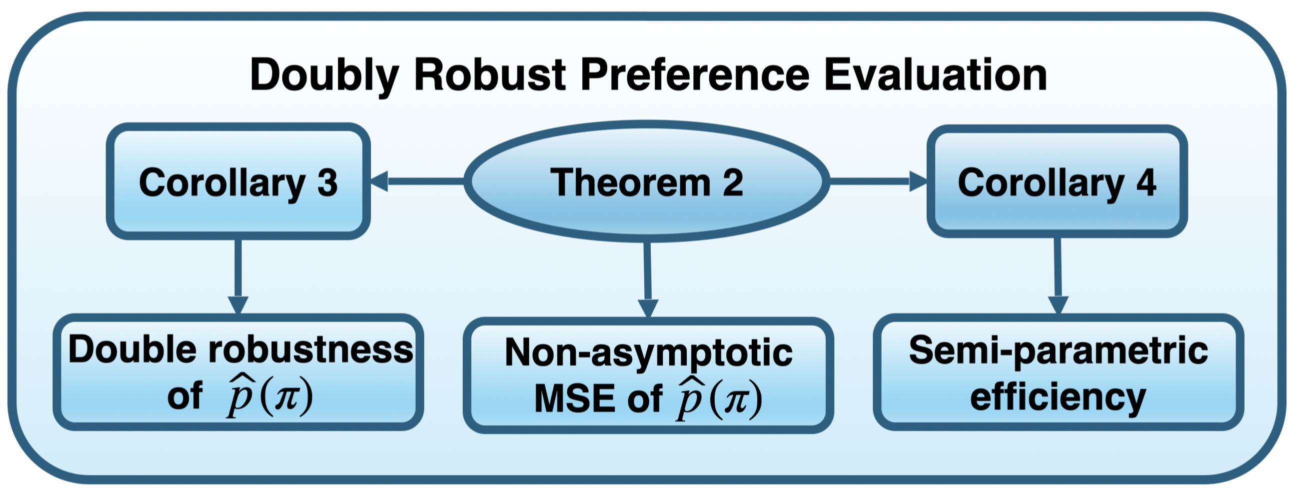
Doubly Robust Preference Evaluation

Goal: Estimate total preference $p^*(\pi) = \mathbb{E}_{y\sim\pi, y'\sim\pi_{\text{ref}}} g^*(X, y, y')$

Our DR Estimator:

$$\widehat{p}(\pi) = \frac{1}{2}\mathbb{E}_{(X,Y^{(1)},Y^{(2)},Z)\sim\mathcal{D}}\Bigg\{\sum_{a=1}^{2}\mathbb{E}_{y\sim\pi(\cdot|X)}[\widehat{g}(X,y,Y^{(a)})] + \sum_{a=1}^{2}(-1)^{a-1}\frac{\pi(Y^{(a)}|X)}{\widehat{\pi}_{\text{ref}}(Y^{(a)}|X)}[Z-\widehat{g}(X,Y^{(1)},Y^{(2)})]\Bigg\}$$

Main Theoretical Results:



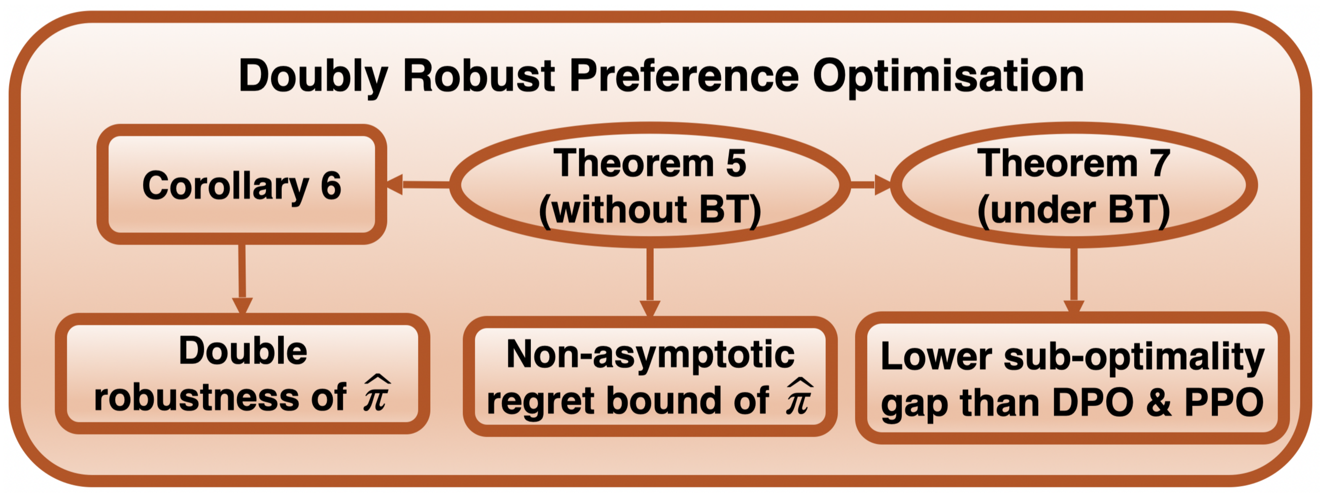
- **Double Robustness:** MSE $\to 0$ when either $\widehat{\pi}_{\text{ref}}$ or $\widehat{g}$ is correct
- **Semiparametric Efficiency:** Achieves smallest-possible MSE when both correct

Doubly Robust Preference Optimization

For Preference Optimization:

$$\widehat{\pi} = \arg\max_{\pi} \widehat{p}(\pi) - \beta \mathrm{KL}(\pi, \widehat{\pi}_{ref})$$

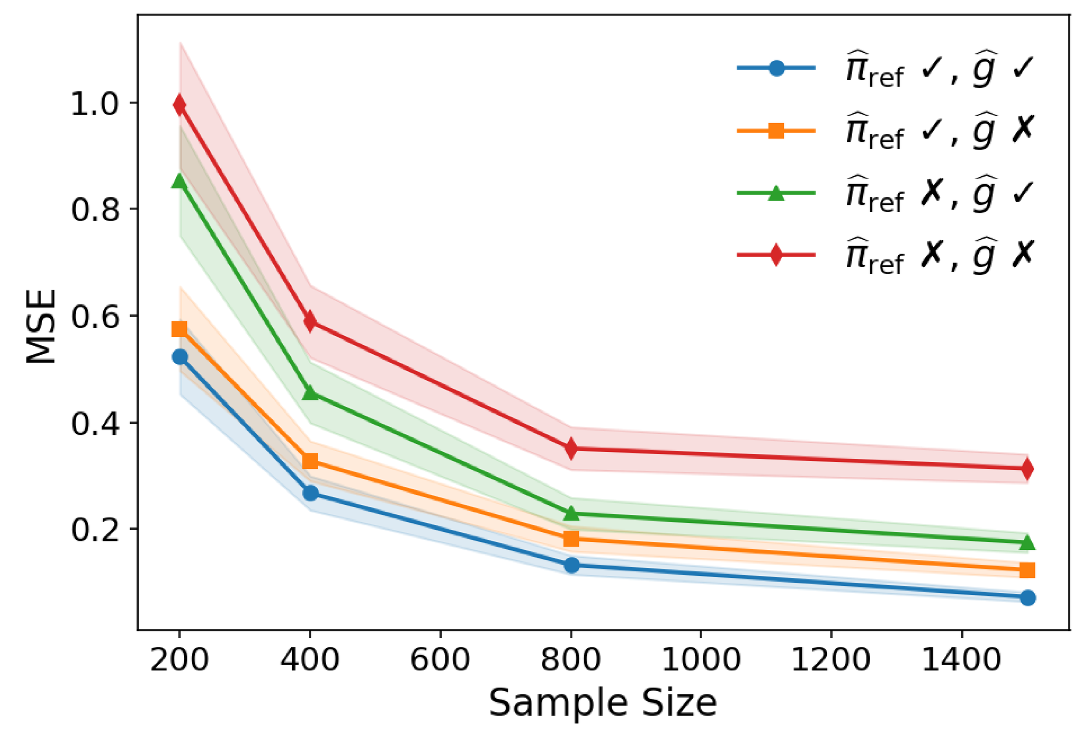
Main Theoretical Results:



- **Double robustness**: Regret of $\widehat{\pi}$ decays to zero when *either* reference policy *or* preference model (not necessarily both) is correct
- **Sub-optimality gaps**:
  - PPO: $O(n^{-1/2} + \|\widehat{r} - r\|)$
  - DPO: $O(n^{-1/2} + \|\widehat{\pi}_{ref} - \pi_{ref}\|)$
  - DRPO: $O(n^{-1/2} + \|\widehat{r} - r\|\|\widehat{\pi}_{ref} - \pi_{ref}\|)$

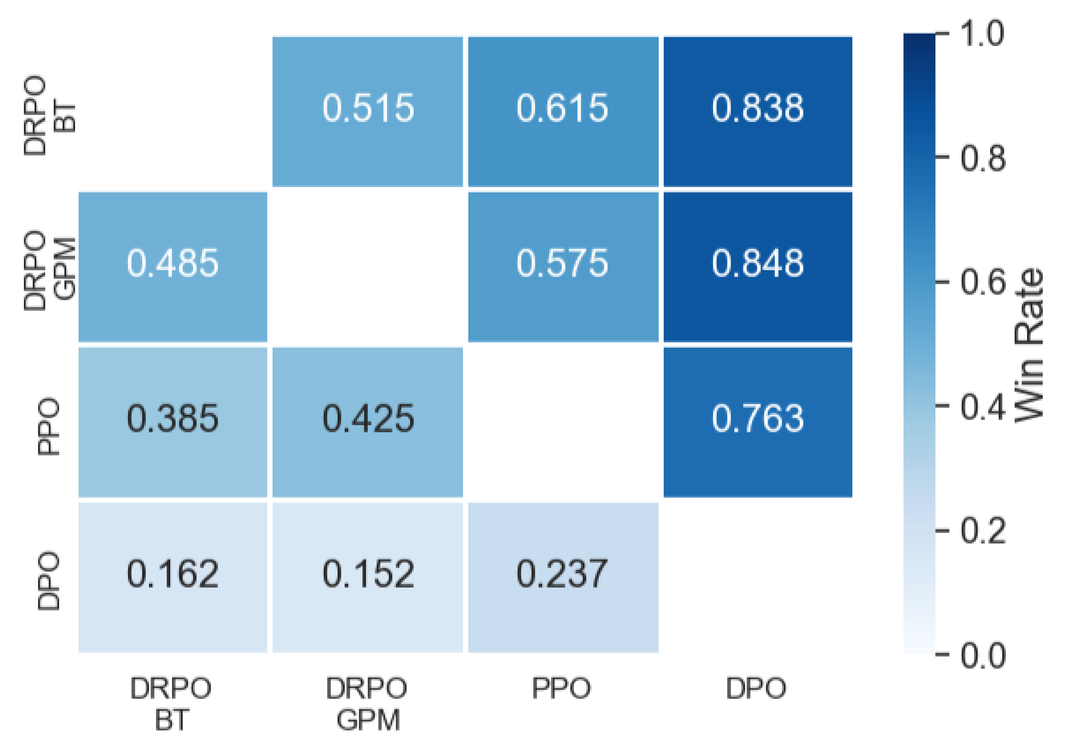
Empirical Results

Applications to IMDb Dataset:

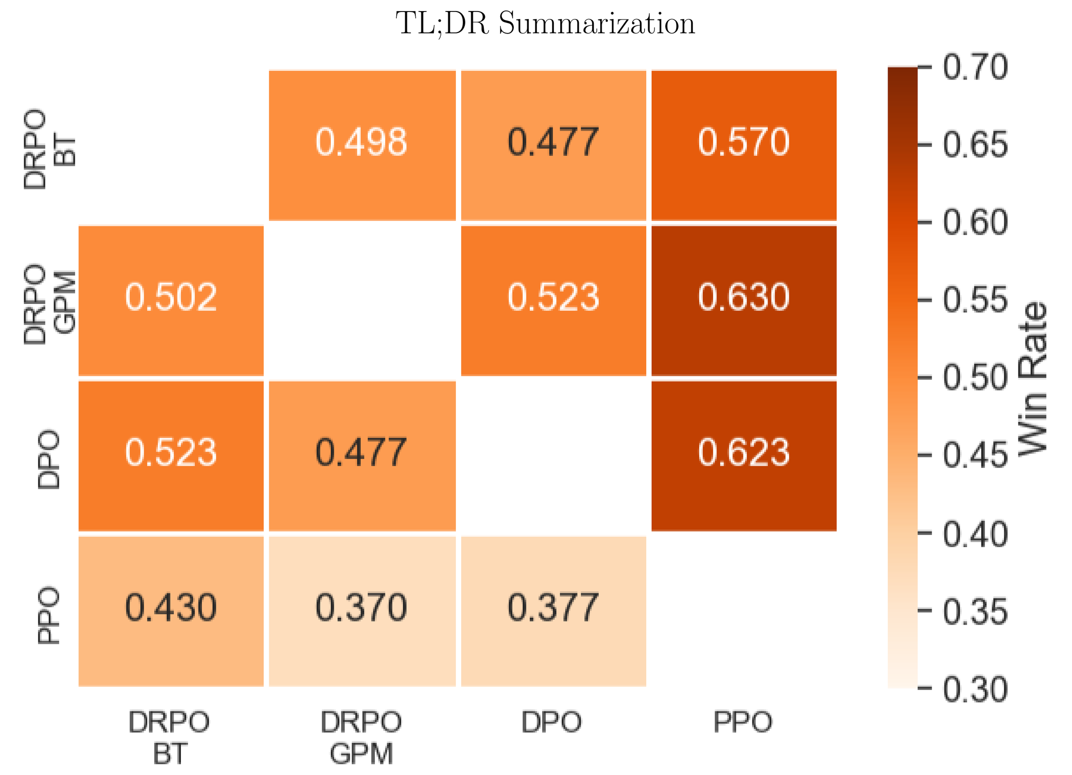


- Evaluate total preference of DPO-trained policy over SFT reference
- Ground truth: 0.681 (computed via Monte Carlo)
- MSE converges to zero when either model is correct

Applications to TL;DR and HH Datasets:



TL;DR Summarization



HH Dialogue

- DRPO (both BT and General Preference Model variants) achieves more robust and often superior performance to PPO and DPO under GPT-4o-mini evaluation
- Robust performance without extensive hyperparameter tuning