Robust Reinforcement Learning from Human Feedback for Large Language Models Fine-Tuning

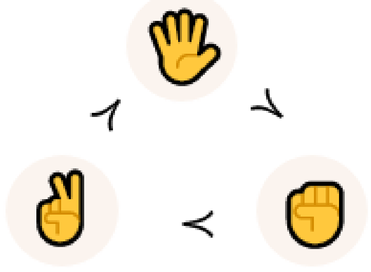
Kai Ye[1] Hongyi Zhou[2] Jin Zhu[1] Francesco Quinzan[3] Chengchun Shi[1]
[1]LSE [2]Tsinghua University [3]University of Oxford

Abstract

Reinforcement learning from human feedback (RLHF) has emerged as a key technique for aligning the output of large language models (LLMs) with human preferences. To learn the reward function, most existing RLHF algorithms use the Bradley-Terry model, which relies on assumptions about human preferences that may not reflect the complexity and variability of real-world judgments. In this paper, we propose a robust algorithm to enhance the performance of existing approaches under such reward model misspecifications. Theoretically, our algorithm reduces the variance of reward and policy estimators, leading to improved regret bounds. Empirical evaluations on LLM benchmark datasets demonstrate that the proposed algorithm consistently outperforms existing methods, with 77-81% of responses being favored over baselines on the Anthropic Helpful and Harmless dataset.

Problem: preference model misspecification

RLHF algorithms for fine-tuning large language models typically require a human preference model, with the Bradley-Terry (BT) model being the most widely adopted. However, it relies on a reward-based preference assumption that imposes several idealized conditions on human judgments:



- **Transitivity**: human preferences are logically ordered ($A \succ B \succ C$ implies $A \succ C$).
- **Context-independence**: preferences between two responses are based solely on the prompts and responses themselves.
- **Perfect rationality**: users provide consistent and deterministic feedback.

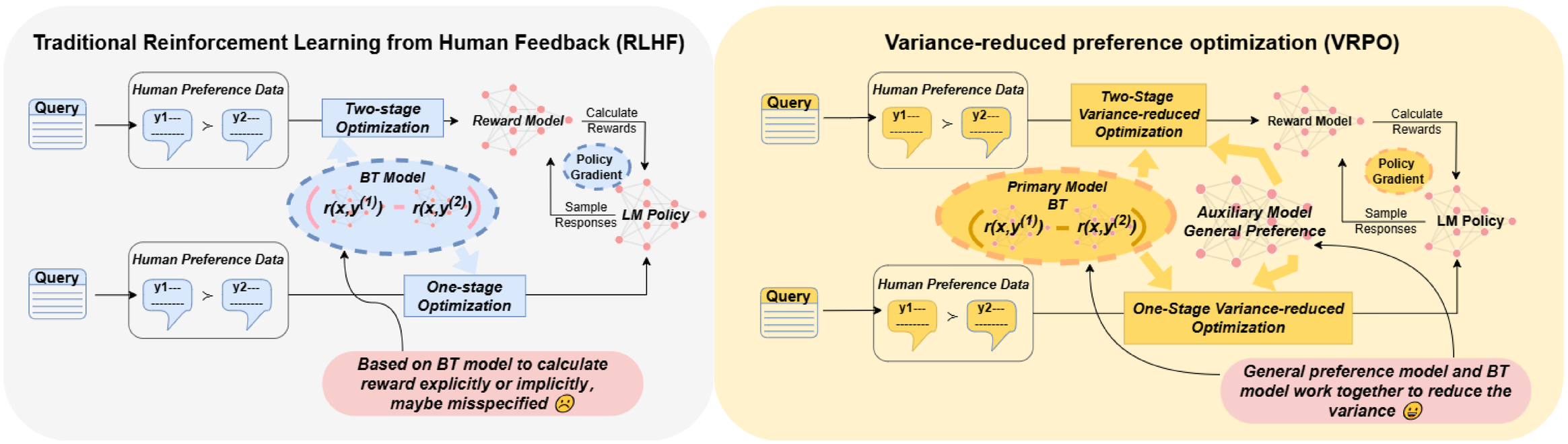
Those assumptions may be misspecified or unrealistic in practice.

Our solution: VRPO

- A flexible pipeline applicable to a variety of existing RLHF algorithms to enhance their sample efficiency under human preference model misspecification.
- Estimate two models for the preference function p^*: (i) A primary, simpler reward-based model p_θ, as in existing algorithms; and (ii) an auxiliary, more complex preference model p_η, designed to enhance the accuracy of the primary model.
- Loss function:

$$\widetilde{\mathcal{L}}(\theta) = \mathbb{E}_n\Big[\ell(X, Y^{(1)}, Y^{(2)}, Z; \theta) - \sum_{u=0}^{1} \ell(X, Y^{(1)}, Y^{(2)}, u; \theta)) p_\eta(X, Y^{(1)}, Y^{(2)}, u) + \sum_{u=0}^{1} \mathbb{E}_{y^{(1)*}, y^{(2)*} \sim \pi_{\text{ref}}(\cdot \mid X)} \ell(X, y^{(1)*}, y^{(2)*}, u; \theta) p_\eta(X, y^{(1)*}, y^{(2)*}, u)\Big].$$

Variance-reduced preference optimization (VRPO) pipeline



VRPO incorporates an auxiliary preference model to reduce the variance of the estimated primary model. Left: The classic one-stage and two-stage optimization schemes in RLHF. Both approaches require fitting a reward model, either explicitly or implicitly, which may lead to model misspecification. **Right:** In contrast, VRPO employs an auxiliary reward-free preference model to better capture human preferences. It works jointly with the primary model for variance reduction and policy improvement.

Theoretical Guarantees

Model Setting	Variance of Estimator	MSE of Estimator	Suboptimality Gap
Misspecified	↓	↓	↓
Correctly specified	→	→	→

Double Robustness

In the correctly specified setting, the target parameter $\bar{\theta} = \arg\min_\theta \mathbb{E}[\widetilde{\mathcal{L}}(\theta)]$, when either the reference policy π_{ref} or the auxiliary preference model p_η is correctly specified.

Variance and MSE reductions

regardless of whether the model is correctly specified or misspecified, we have

$$\|\mathbb{E}(\widehat{\theta}) - \bar{\theta}\|_2 = O\Big(\frac{d}{n\lambda_{\min}}\Big), \qquad \|\mathbb{E}(\widetilde{\theta}) - \bar{\theta}\|_2 = O\Big(\frac{d}{n\lambda_{\min}}\Big),$$

and

$$\text{Var}(\widehat{\theta}) - \text{Var}(\widetilde{\theta}) = \underbrace{\frac{1}{n} A^{-1}(\bar{\theta}) H A^{-1}(\bar{\theta})}_{\text{variance reduction}} + \underbrace{O\left(\frac{d^{3/2}}{n^{3/2}\lambda_{\min}^2}\right)}_{\text{nonlinear residual}} + \underbrace{O\left(\frac{\|p_\eta - p^*\|_\infty^2}{n\lambda_{\min}^2}\right)}_{\text{model specification error}},$$

where H denotes certain positive semi-define matrix, $A(\bar{\theta}) := -\mathbb{E}\left\{\frac{\partial^2}{\partial\theta^2}\mathcal{L}(y, A, x; \bar{\theta})\right\}$, $\lambda_{\min}$ denote the minimum eigenvalue of $A(\bar{\theta})$ and $\|p_\eta - p^*\|_\infty$ denotes the difference between p_η and p^* in supremum norm.

Variance and MSE reductions

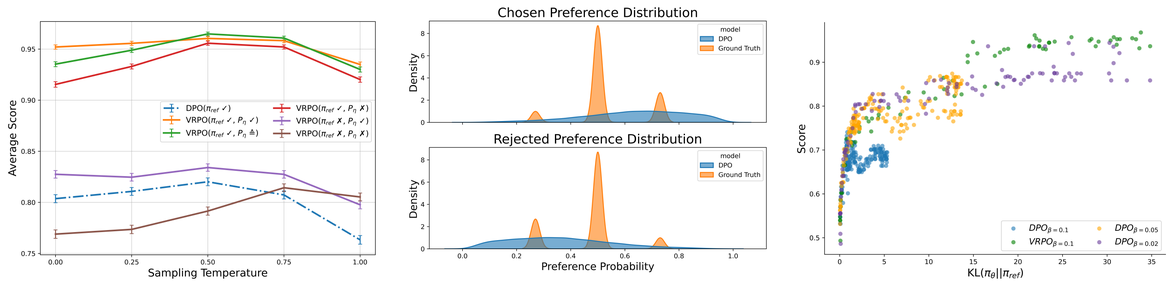
regardless of whether the model is correctly specified or misspecified, we have

$$\mathbb{E}\left\{R(\widehat{\theta})\right\} = R(\bar{\theta}) + \text{trace}\left(\text{Var}(\widehat{\theta})(-\nabla_\theta^2 J(\bar{\theta}))\right) + O\left(\frac{d^{3/2}}{n^{3/2}\lambda_{\min}^3}\right),$$

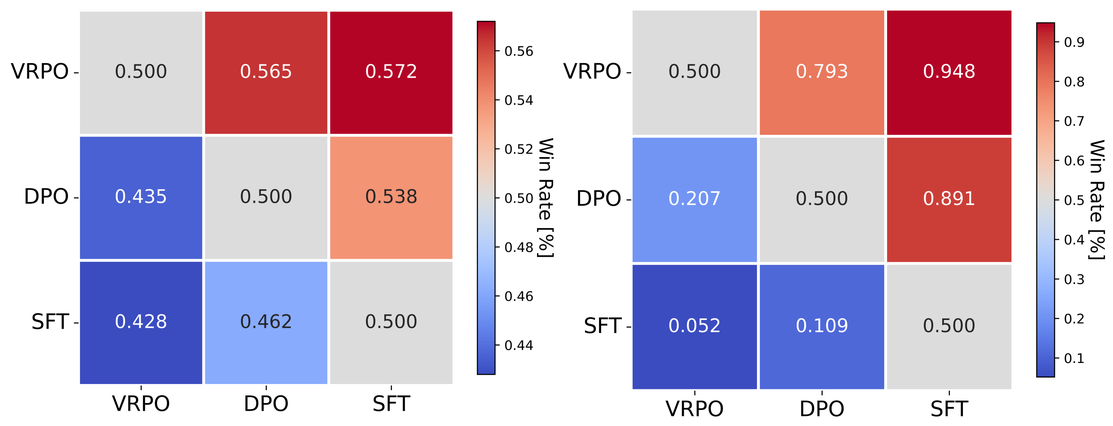
$$\mathbb{E}\left\{R(\widetilde{\theta})\right\} = R(\bar{\theta}) + \text{trace}\left(\text{Var}(\widetilde{\theta})(-\nabla_\theta^2 J(\bar{\theta}))\right) + O\left(\frac{d^{3/2}}{n^{3/2}\lambda_{\min}^3}\right),$$

where $\text{Var}(\widehat{\theta})$ and $\text{Var}(\widetilde{\theta})$ denote the covariance matrices of $\widehat{\theta}$ and $\widetilde{\theta}$, respectively.

Experiments performance



Comparisons in IMDb dataset. **Left** panel represents the expected reward in different VRPO setting compared to DPO, for example (π_{ref} ✓, P_η ✗) means the reference model is correctly specified and the preference model is misspecified, and $P_\eta \hat{=}$ means the preference model is estimated, demonstrating the robustness of our method. **Middle** plane illustrates the difference in preference probability distributions between the ground truth and the DPO estimation for both the Chosen and Rejected responses. **Right** The panel reports the expected reward versus KL-divergence for VRPO and DPO, demonstrating the quality of the optimization.



Head-to-head comparisons between VRPO, DPO, SFT. Win rates are evaluated by GPT-4o-mini. **Left** panel displays the win rate in summarization task. In both tasks, VPRO outperforms DPO by defeating it directly and demonstrating a higher win rate against SFT. **Right** panel displays the win rate in Anthropic-HH one-step dialogue task.