A Minimaximalist Approach to Reinforcement Learning from Human Feedback

Gokul Swamy [1] **Christoph Dann** [2] **Rahul Kidambi** [2] **Zhiwei Steven Wu** [1] **Alekh Agarwal** [2]

ICML'24, Google Research, 84 citation

Background

Bradley-Terry Reward Model

$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp\left(r^*(x, y_1)\right)}{\exp\left(r^*(x, y_1)\right) + \exp\left(r^*(x, y_2)\right)}.$$

Learning Reward Model

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))\right]$$

Learning Policy (PPO, TRPO)

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y|x)} \left[r_\phi(x, y)\right] - \beta \mathbb{D}_{\mathrm{KL}} \left[\pi_\theta(y \mid x) \,||\, \pi_{\mathrm{ref}}(y \mid x)\right]$$

Learning Policy (DPO, SPO)

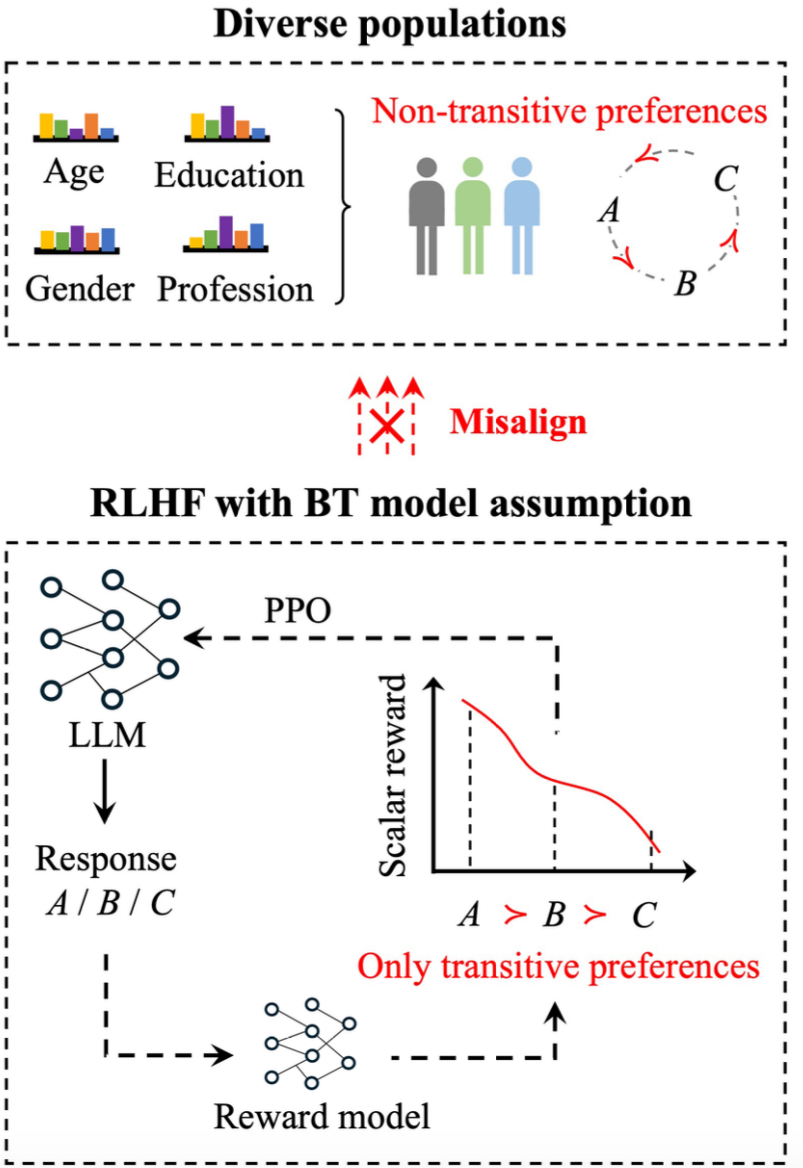
$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta; \pi_{\mathrm{ref}}) = -\mathbb{E}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\mathrm{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\mathrm{ref}}(y_l \mid x)}\right)\right]$$

$$\mathcal{L}_{\mathrm{SimPO}}(\pi_\theta) = -\mathbb{E}\left[\log \sigma\left(\frac{\beta}{|y_w|} \log \pi_\theta(y_w \mid x) - \frac{\beta}{|y_l|} \log \pi_\theta(y_l \mid x) - \gamma\right)\right]$$

Problem

- Non-transitive Preferences
  - Conventional RLHF methods typically rely on the Bradley-Terry (BT) assumption for preference modeling, which presumes **transitivity in human preferences**—if response *A* is preferred over *B*, and *B* over *C*, then *A* should also be preferred over *C*.

$$A \succ B, B \succ C \Rightarrow A \succ C$$



[1] Magnetic Preference Optimization: Achieving Last-iterate Convergence for Language Model Alignment. ICLR'25

Problem

- The Limitation of ELO score
  - ELO score fails to capture the correct preference ordering between policies, even in **transitive situations**.

$$\mathbb{P}(i \text{ beats } j) = \sigma(\alpha(\mathrm{u}_i - \mathrm{u}_j))$$

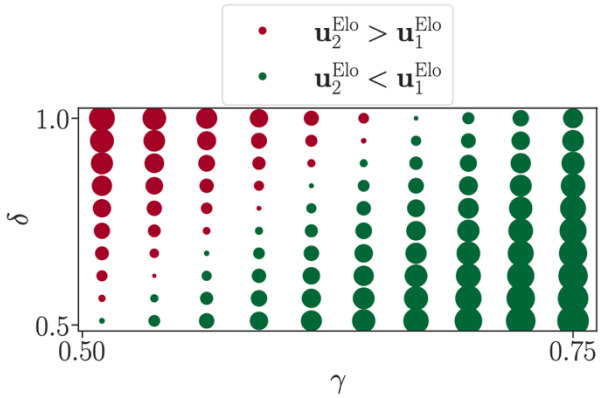


Figure 2: **Elo Score Fails to Rank Players for Some Transitive Games.** This figure displays the difference $\mathrm{u}_1^{\mathrm{Elo}} - \mathrm{u}_2^{\mathrm{Elo}}$ between the Elo scores of $\mathrm{u}_1^{\mathrm{Elo}}$ of player 1, and $\mathrm{u}_2^{\mathrm{Elo}}$ of player 2 (computed using the stationary Elo score in Equation (1)). The difference of the Elo scores is displayed for multiple probability matrices $\mathbf{P}^{(\gamma,\delta)}$ of the transitive game (Example 3). Red dots indicate that $\mathrm{u}_2^{\mathrm{Elo}} > \mathrm{u}_1^{\mathrm{Elo}}$ and green dots indicate $\mathrm{u}_1^{\mathrm{Elo}} > \mathrm{u}_2^{\mathrm{Elo}}$, the size of the dots is proportional to $|\mathrm{u}_1^{\mathrm{Elo}} - \mathrm{u}_2^{\mathrm{Elo}}|$.

Example 3. *Here we define a family of three-player transitive games for all $\gamma, \delta \in (0.5, 1]$. Contrary to Elo games (Example 1), outcome probabilities might be non-additive:*

$$\mathbf{P}^{(\gamma,\delta)} = \begin{pmatrix} 0.5 & \gamma & \gamma \\ 1-\gamma & 0.5 & \delta \\ 1-\gamma & 1-\delta & 0.5 \end{pmatrix} .$$

Example 3 describes the payoff matrix of a game that is transitive for $\gamma, \delta \in (0.5, 1]$, however when γ is close to 0.5 and δ is close to 1—i.e., the second player slightly loses against the first one and significantly wins against the third one—Elo score fails to assign scores which yield correct matchup predictions between players. Figure 2 displays the set of values γ, δ for which the Elo score fails (in red) and succeeds (in green) to correctly estimate the probability of winning between the first and the second players of the game $\mathbf{P}^{(\gamma,\delta)}$ (Example 3). Despite the game being transitive (player 1 beats player 2 and 3 and player 2 beats player 3), there exists a significant range of values γ, δ, for which the Elo score assigns a larger score to player 2 than player 1, and thus wrongly predicts the outcome of the confrontation.

[1] On the Limitations of the Elo, Real-World Games are Transitive, not Additive. AISTATS'23

Problem

- The Diversity of Human Preference
  - Given the inherent stochasticity of human preferences, one often learns a reward model that leads to a **collapse in generation diversity**.
  - Due to either finite sample or optimization error, we can easily learn a model that assigns a **slightly higher reward** to one option over the other. Then, if we were to optimize our policy under this model, we would learn to (almost) exclusively select one option, **leaving half of the population unsatisfied**.

[1] Nash Learning from Human Feedback. ICML'24

Motivation

Intuitively, this means that while **we don't always make everyone happy** (an impossibility), we never pick a solution that makes a significant portion of the population consistently unhappy.

纳什均衡

假设有两个小偷A和B联合犯事、私入民宅被警察抓住。警方将两人分别置于不同的两个房间内进行审讯，对每一个犯罪嫌疑人，警方给出的政策是：如果一个犯罪嫌疑人坦白了罪行，交出了赃物，于是证据确凿，两人都被判有罪。如果另一个犯罪嫌疑人也作了坦白，则两人各被判刑8年；如果另一个犯罪嫌人没有坦白而是抵赖，则以妨碍公务罪（因已有证据表明其有罪）再加刑2年，而坦白者有功被减刑8年，立即释放。如果两人都抵赖，则警方因证据不足不能判两人的偷窃罪，但可以私入民宅的罪名将两人各判入狱1年。

此时产生了两个嫌疑人之间的一场博弈：

A/B	坦白	抵赖
坦白	-8，-8	0，10
抵赖	-10，0	-1，-1

Define

Preference Oracle. In the preference-based RL setup, we are given query access to a *preference function*

$$\mathcal{P} : \Xi \times \Xi \to [-1, 1] \tag{1}$$

which, given two trajectories $\xi_1, \xi_2 \in \Xi \times \Xi$, outputs a scalar that indicates the preferred trajectory. Explicitly, given some comparison function $P(\xi_1 \succ \xi_2)$, we define $\mathcal{P}(\xi_1, \xi_2) = 2P(\xi_1 \succ \xi_2) - 1$. Practically, this could be

By construction, preference functions are anti-symmetric, i.e. $\forall \xi_1, \xi_2 \in \Xi \times \Xi$, $\mathcal{P}(\xi_1, \xi_2) = -\mathcal{P}(\xi_2, \xi_1)$. Similarly, we have that $\forall \xi \in \Xi$, $\mathcal{P}(\xi, \xi) = 0$.

We assume access to a convex and compact policy class $\Pi \subseteq \{\mathcal{S} \to \Delta(\mathcal{A})\}$. With a slight abuse of notation, we can now define the preference function over policy pairs as

$$\mathcal{P}(\pi_1, \pi_2) \triangleq \mathbb{E}_{\xi_1 \sim \pi_1, \xi_2 \sim \pi_2}[\mathcal{P}(\xi_1, \xi_2)]. \tag{2}$$

Social Choice Theory

Given choices from a population of raters that are represented as a preference function $\mathcal{P}$, social choice theory (Sen, 1986) studies the question of how best to select options that satisfy the diversity of preferences inherent in the said population. For example, consider the set of preferences $\mathcal{P}_1$ over options (a, b, c, d) in Figure 3.

	a	b	c	d
a	0	+1	+1	−1
b	−1	0	+1	−1
c	−1	−1	0	+1
d	+1	+1	−1	0

Figure 3: An intransitive preference function $\mathcal{P}_1$ over (a, b, c, d). $\mathcal{P}_1(x, y) = 1$ if $P(x \succ y) = 1$, -1 if $P(x \succ y) = 0$, and 0 if $P(x \succ y) = 0.5$. Observe that there is no unique Copeland Winner.

Given this preference function, perhaps the most natural idea would be to pick the option that beats the largest number of other options. In the above matrix, this would be either option a or d as they have the largest row sums. More formally, this technique is known as a *Copeland Winner* and can be expressed mathematically as

$$\mathsf{CW}(\mathcal{P}) \triangleq \operatorname*{argmax}_{\pi \in \Pi} \sum_{\pi' \in \Pi} \mathcal{P}(\pi, \pi'). \tag{3}$$

While intuitively appealing, Copeland Winners are often not unique as in our above example, raising the question of how to break ties. For example, if half of the group feels like $a \succ d$ and the other half like $d \succ a$, picking either option would leave half of the group unsatisfied. This problem only gets worse as the number of options to choose between increases, as there is unlikely to be a single option that *everyone* prefers to *every* other option (Dudík et al., 2015).[3]

In essence, approaches that train reward models like reward-based RLHF (or implicitly assume them like DPO) are akin to computing Copeland Winners. Observe that our above matrix has an *intransitivity*: $a \succ c, c \succ d, d \succ a$. This means that *no* reward function can explain the above preferences as it would need to satisfy $r(a) > r(c), r(c) > r(d)$ and $r(d) > r(a)$ simultaneously, an impossibility. Thus, the

MiniMax Winner

In this paper we will consider the objective of finding a policy π^* which is preferred over any other alternative policy:

$$\pi^* \stackrel{\text{def}}{=} \arg\max_{\pi} \min_{\pi'} \mathcal{P}(\pi \succ \pi') \,. \tag{1}$$

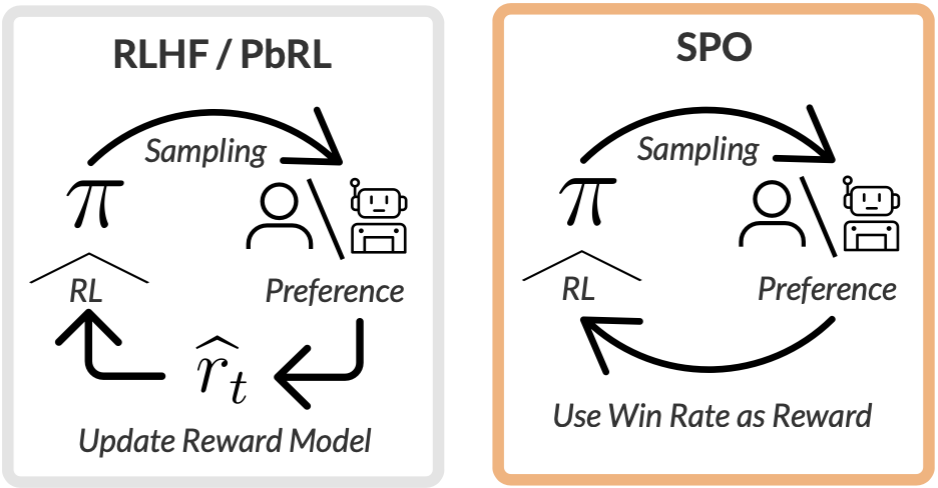
This objective implicitly defines a two-player game, in which the players select policies π and π', the first player receiving a payoff of $\mathcal{P}(\pi \succ \pi')$, and the second player receiving $\mathcal{P}(\pi' \succ \pi) = 1 - \mathcal{P}(\pi \succ \pi')$. This is therefore a two-player, antisymmetric, constant-sum game, and it follows that when both players use a policy π^* solving Equation (1), this is a *Nash equilibrium* for this game, by the minimax theorem (von Neumann, 1928). This is the fundamental solution concept we study in this paper.

Self-Play Preference Optimization (SPO)

Common algorithms like gradient descent satisfy this property (Zinkevich, 2003). See Hazan et al. (2016) for a more extensive list. We define the **SPO Loss** at round $t \in [T]$ as the negative preference against the current iterate p_t,

$$\ell_t^{\mathtt{SPO}}(p) \triangleq \mathbb{E}_{\pi \sim p, \pi' \sim p_t}[-\mathcal{P}(\pi, \pi'))]. \tag{6}$$

Queryable Preference Oracle



Self-Play Preference Optimization (SPO)

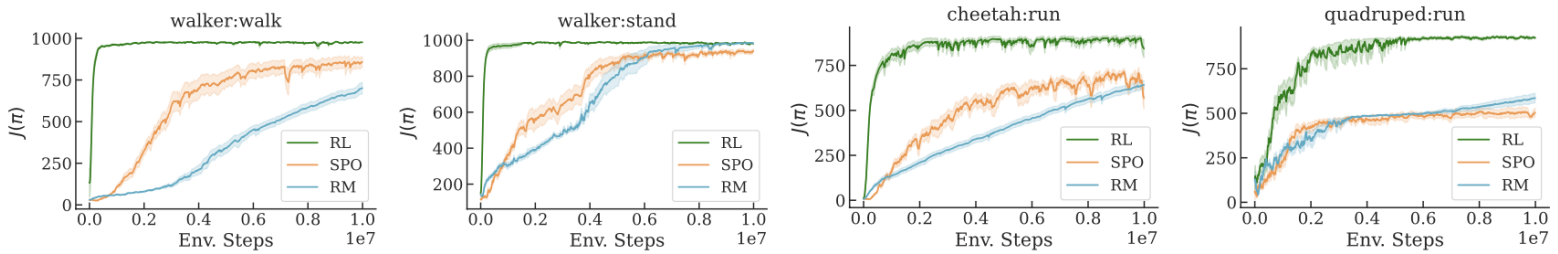
Algorithm 2 SPO (Practical Version)

Input: Iterations T, Preference fn. $\mathcal{P}$, Queue size B, Reinforcement learning algo. RL : $\Pi \times \mathcal{D} \to \Pi$.
Output: Trained policy π.
Initialize $\pi_1 \in \Pi$, Queue $\mathcal{Q} \leftarrow [\xi_{1:B} \sim \pi_1]$.
for t in $1 \dots T$ **do**
  Sample $\xi_t \sim \pi_t$.
  `// Win-rate over queue as reward`
  Compute $r_t(\xi_t) = \frac{1}{|\mathcal{Q}|} \sum_{q=1}^{B} \mathcal{P}(\xi_t, \xi_q)$.
  Set $r_t^h = r_t(\xi_t)/H, \forall h \in [H]$.
  `// use PPO, TRPO, SAC ...`
  $\pi_{t+1} \leftarrow \mathrm{RL}(\pi_t, \mathcal{D} = \{(s_t^h, a_t^h, r_t^h)\}_{h \in [H]})$.
  $\mathcal{Q} \leftarrow [\xi_2, \dots, \xi_B, \xi_t]$.
end for
Return best of $\pi_{1:T}$ on validation data.

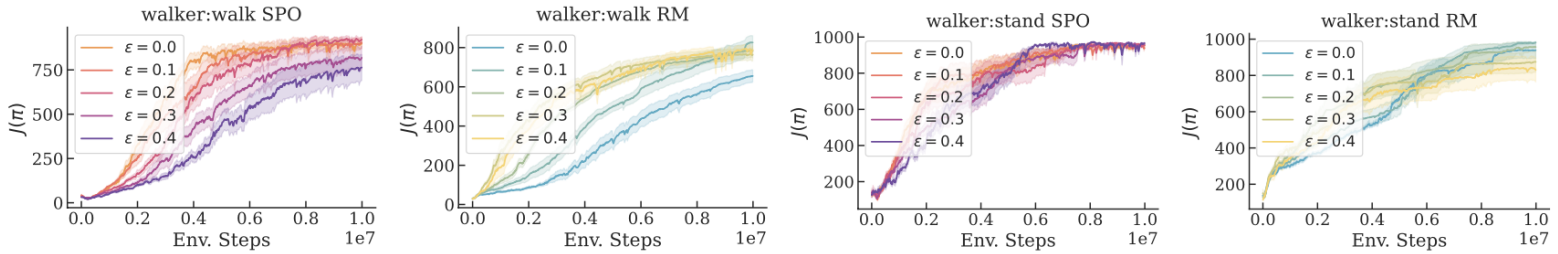
Approach	Example	Compounding Errors	Intransitive Prefs	Learning Setup
Offline, Reward-Based	DPO (Rafailov et al., 2023)	✗, Example D.1	✗, Theorem 2.4	Offline, log-loss
Online, Reward-Based	PPO (Ouyang et al., 2022)	✓	✗, Theorem 2.4	Online RL
Online, Dueling	DBGD (Yue & Joachims, 2009)	✓	✓	Online *adver.* RL
Online, Preference-Based	**SPO** (ours)	✓	✓	Online RL

Table 1: An taxonomy of RLHF algorithms and the sorts of issues they are robust to.

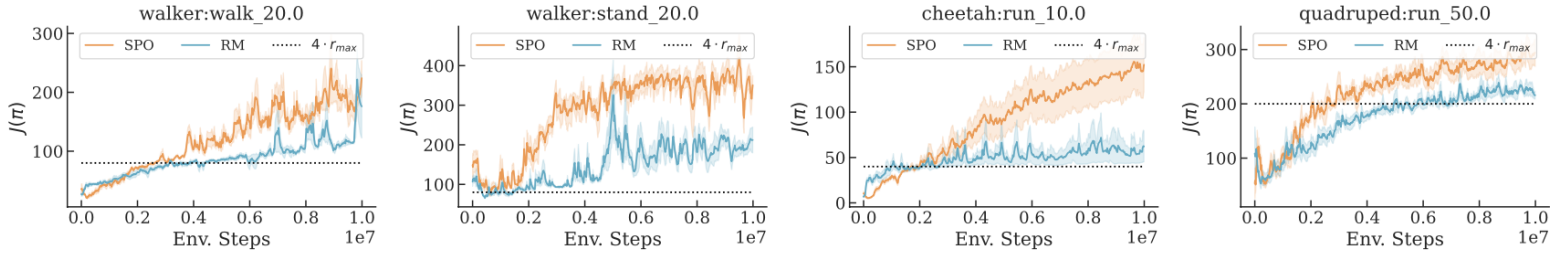
Experiments



(a) Maximum Reward Preferences

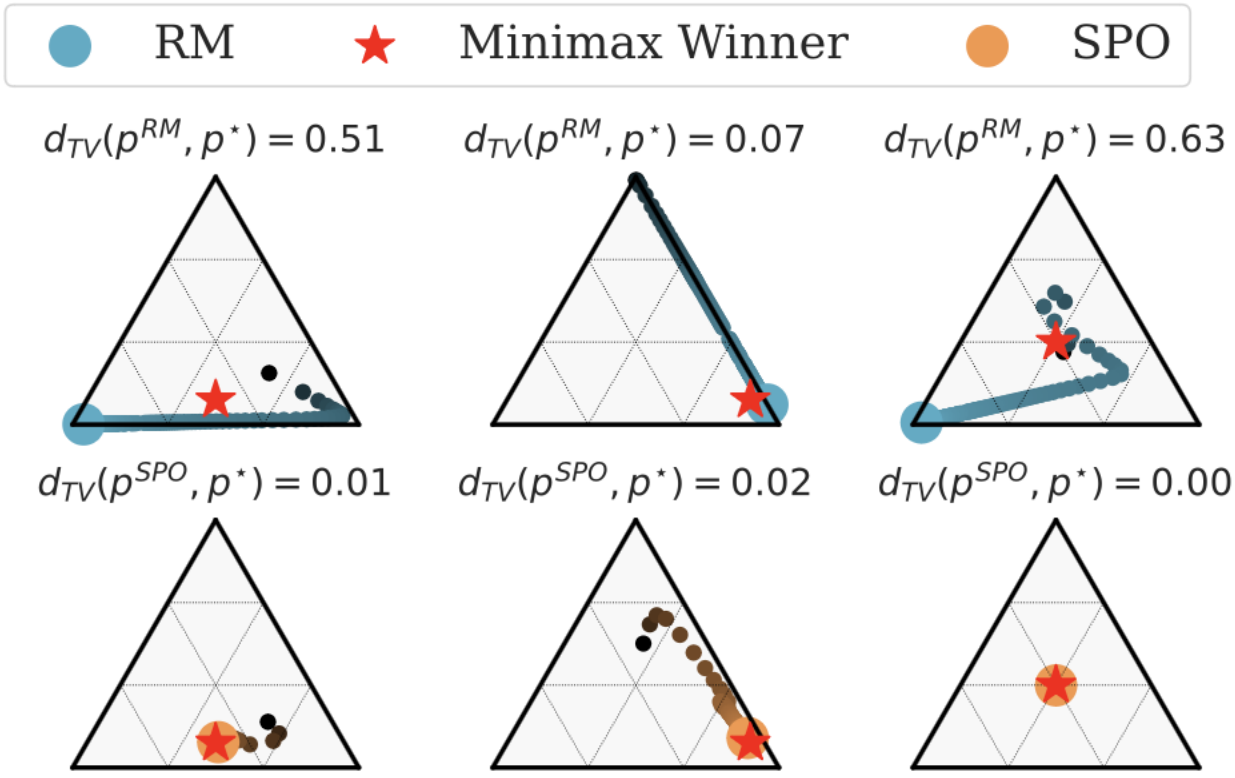


(b) Stochastic Preferences



(c) Non-Markovian Preferences

Experiments



(a) Intransitive Prefs. (Discrete)

Discussion

[1] Magnetic Preference Optimization: Achieving Last-iterate Convergence for Language Model Alignment. ICLR'25, Peking University

[2] Nash Learning from Human Feedback. ICML'24, Google DeepMind

[3] A Minimaximalist Approach to Reinforcement Learning from Human Feedback. ICML'24, Google Reasearch.

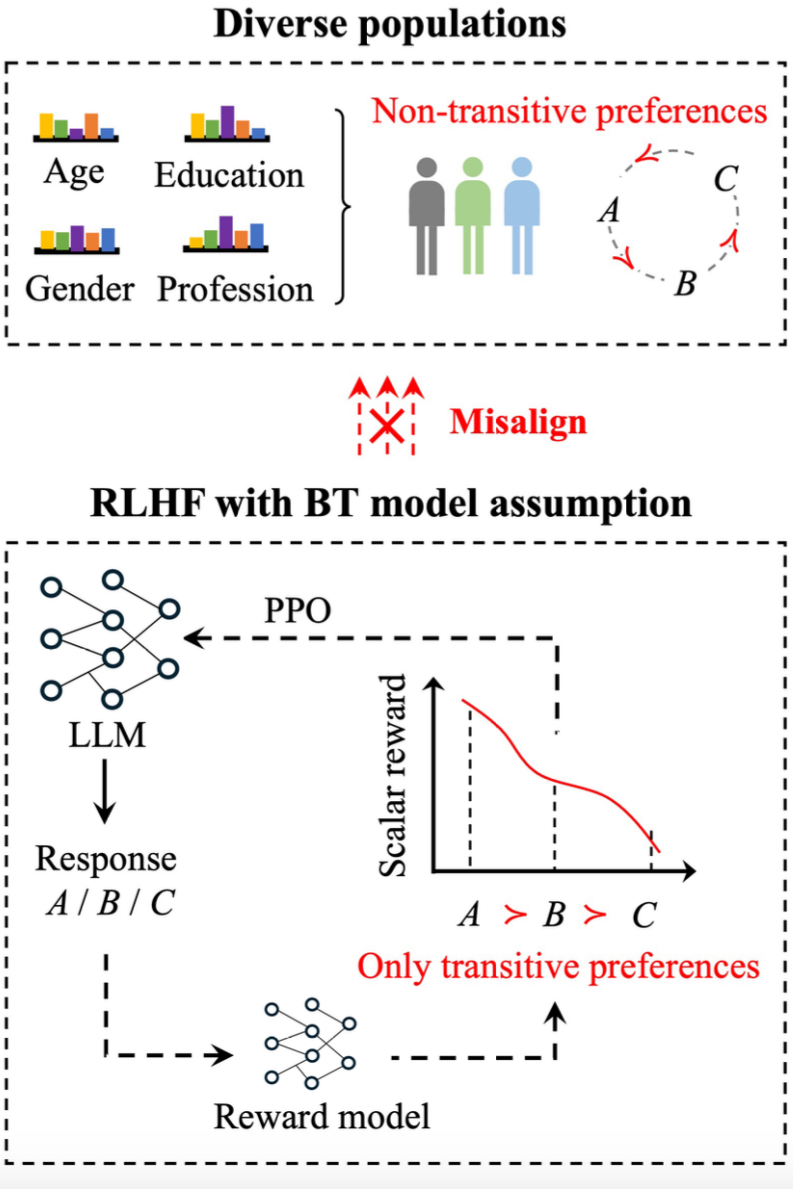
[4] On the Limitations of the Elo, Real-World Games are Transitive, not Additive. AISTATS'23, CIFAR.

Nash equilibrium in evaluating LLMs/VLMs?

Discussion

EM + DPO (Reward Model)?

human preference通常不具备传递性，这让学习reward函数十分困难，reward model很难捕捉到不同case下的偏好。有文章认为人群分布的多样性决定了偏好的非传递性，确实很自然，**比如专业的和非专业的看同一个东西关注点确实不同，偏好自然不同**，所谓外行看热闹内行看门道。拍脑袋一个想法是，**相同batch（分布）下的人类偏好是否具备传递性**，如果这个假设成立，我们其实可以构造一些隐藏变量去学习不同分布的聚类关系，用EM算法来优化，是不是合理



Thanks