Strong Preferences Affect the Robustness of Preference Models and Value Alignment

Ziwei Xu   Mohan Kankanhalli

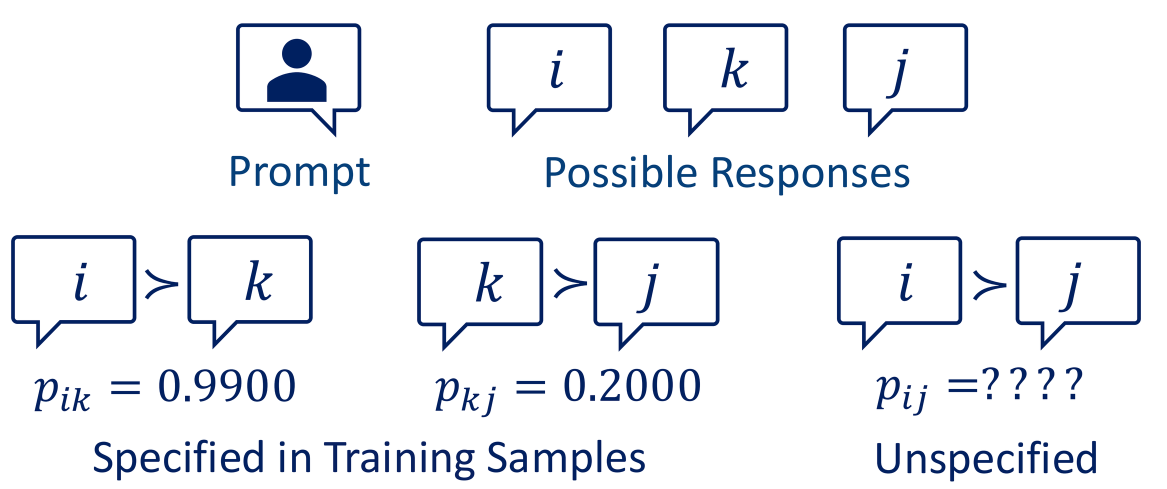
NUS National University of Singapore | Artificial Intelligence Institute   ICLR

TL; DR

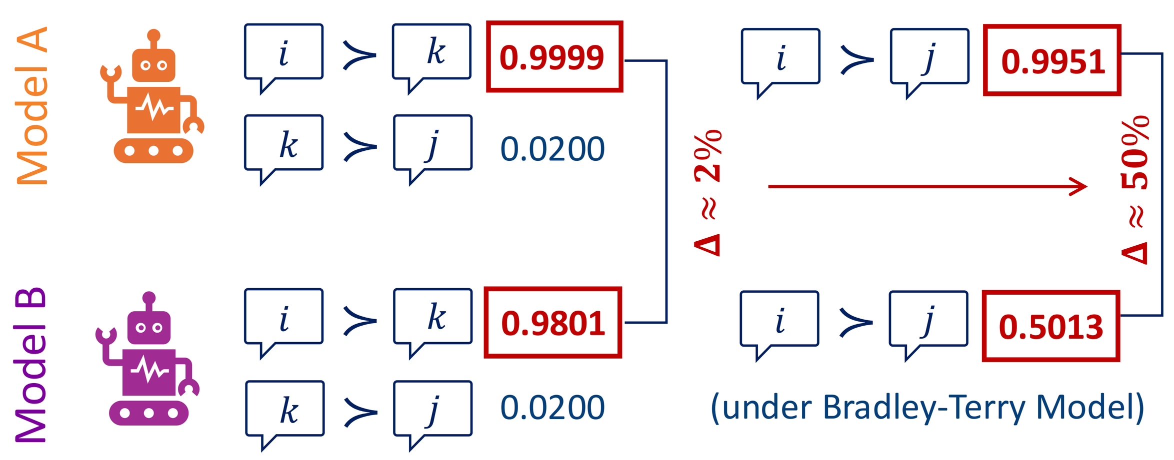
- We study the robustness of value alignment by analyzing the sensitivity of preference models, a core component of value alignment.
- We show that under the Plackett-Luce model, preference probabilities can change significantly due to small changes in the learned preference distribution.
- We characterize this sensitivity: it occurs with strong preferences with probabilities close to 0 and 1.

Introduction

We cannot Specify all the Preferences



Models could Learn Slightly Different Preferences



Question We Ask

How sensitive a preference probability is with respect to changes in other preference probabilities?

Assumptions for Pairwise Preference Models

Asm 1: Preference probabilities only depend on score differences
$$p_{ij} = p(y_i \succ y_j) = g(s_i - s_j), g \in \mathbb{R} \to (0,1)$$
Asm 2: $g(x)$ is strictly increasing $\Leftrightarrow (s_i - s_j \uparrow \Leftrightarrow p_{ij} \uparrow)$

Asm 3: $\lim_{x \to -\infty} g(x) = 0, \lim_{x \to +\infty} g(x) = 1$

Asm 4: $\forall x \in \mathbb{R}, g(x) + g(-x) = 1 \Leftrightarrow p_{ij} + p_{ji} = 1$

Asm 5: $g(x)$ is continuously differentiable

Special case: Bradley-Terry model $g_{BT}(x) = \frac{1}{1+\exp(-x)}$

Analysis

Measuring Sensitivity

Consider a multivariable function $h(\boldsymbol{x}) = h(x_1, x_2, \dots, x_L)$.

- M**-sensitivity**: $h(\boldsymbol{x})$ is M-sensitive to x_i at $\boldsymbol{x}'$ if $\left|\frac{\partial h}{\partial x_i}\big|_{\boldsymbol{x}=\boldsymbol{x}'}\right| > M$.
- M**-sensitivity region of** h: $\Omega_M(h, x_i)$: $\left\{\boldsymbol{x}' \in \mathrm{Dom}(h): \left|\frac{\partial h}{\partial x_i}\big|_{\boldsymbol{x}=\boldsymbol{x}'}\right| > M\right\}$.

Analysis for Pairwise Preference Models

Lemma 1 The unspecified p_{ij} is a function of p_{ik} and p_{kj}:
$$p_{ij} = g(s_i - s_j) = g(s_i - s_k + s_k - s_j) = g\left(g^{-1}(p_{ik}) + g^{-1}(p_{kj})\right) = p_{ij}(p_{ik}, p_{kj})$$

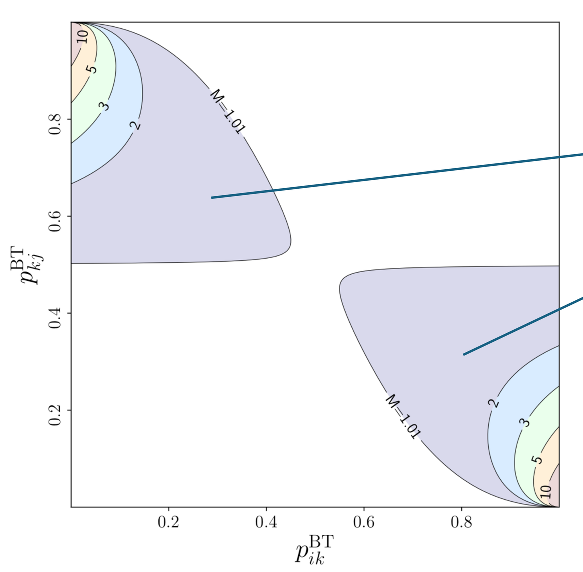
Theorem 1 For all $M > 0$, there exists $0 < p_0, p'_{kj} < 1$, such that $p_{ij}(p_{ik}, p'_{kj})$ is M-sensitive to p_{ik} for all $p_0 < p_{ik} < 1$. Similarly, there exists $0 < p_1, p''_{kj} < 1$, such that $p_{ij}(p_{ik}, p''_{kj})$ is M-sensitive to p_{ik} for all $0 < p_{ik} < p_1$.

- p_{ij} can be arbitrarily sensitive to p_{ik} when p_{ik} (and p_{kj}) are close to 0 or 1.

Characterizing Sensitivity for the Bradley-Terry (B-T) Model

- **When is B-T model sensitive?**

$$p_{ij}^{BT} = 1 \Big/ \left(1 + \frac{(1-p_{ik}^{BT})(1-p_{kj}^{BT})}{p_{ik}^{BT} p_{kj}^{BT}}\right)$$



$$0 < p_{ik}^{BT} < \gamma_0, 1 - \frac{1}{1+M} < p_{kj}^{BT} < 1$$
$$\gamma_0 < p_{ik}^{BT} < 1, 0 < p_{kj}^{BT} < \frac{1}{1+M}$$

$\gamma_0 = \gamma_0(M, p_{kj}^{BT})$ (See Eqn. 5)

$$A\left(\Omega_M(p_{ij}^{BT}, p_{ik}^{BT})\right) = \frac{1}{2}\ln\frac{M-1}{M+1} + \frac{1}{2\sqrt{M}}\ln\frac{\sqrt{M}+1}{\sqrt{M}-1}$$

Area of the region

Back to the example:

$(p_{ik}^A, p_{kj}^A) = (0.9999, 0.0200) \in \Omega_{48}(p_{ij}^{BT}, p_{ik}^{BT})$ $(p_{ik}^B, p_{kj}^B) = (0.9801, 0.0200) \in \Omega_{12}(p_{ij}^{BT}, p_{ik}^{BT})$

Extension to the Plackett-Luce Model

Let $\mathcal{O} = \{o_1, o_2, \dots, o_K\}$ be the set of all the options. Let $\boldsymbol{\omega} = (o_{\omega_1}, o_{\omega_2}, \dots, o_{\omega_K}) \in Perm(\mathcal{O})$ be a preference over the options. Under the Plackett-Luce Model:

Lemma 5 Let $\boldsymbol{\omega}$ be a K-tuple preference. Then $p_{\boldsymbol{\omega}}^{(K)}$, under the Plackett-Luce Model, can be written as a function of $p_{\boldsymbol{\omega}_{uv}}^{(K)} / p_{\boldsymbol{\omega}_{vu}}^{(K)}$, where $1 \le u, v \le K$, $\boldsymbol{\omega}_{uv} = (\boldsymbol{\omega}'_{uv}; \omega_u, \omega_v)$ and $\boldsymbol{\omega}_{vu} = (\boldsymbol{\omega}'_{uv}; \omega_v, \omega_u)$, and $\boldsymbol{\omega}'_{uv}$ is any $(K-2)$-permutation of $\mathcal{O} \backslash \{\omega_u, \omega_v\}$:

$$p_{\boldsymbol{\omega}}^{(K)} = \prod_{u=1}^{K-1} \frac{1}{1 + \sum_{v=u+1}^{K} p_{\boldsymbol{\omega}_{uv}}^{(K)} / p_{\boldsymbol{\omega}_{vu}}^{(K)}}$$

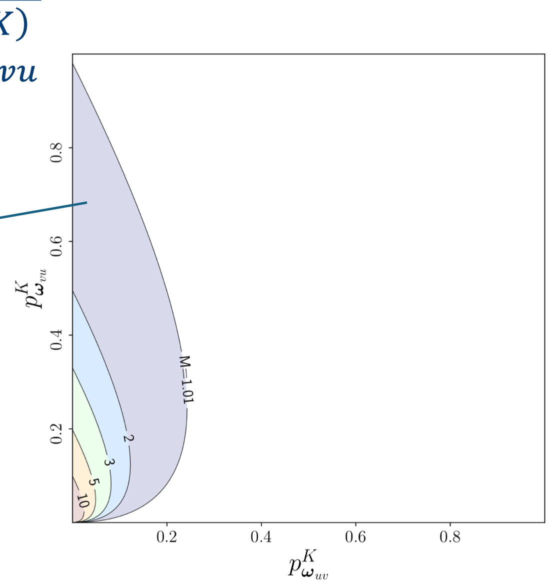
- **When is P-L model sensitive?**

$$0 < p_{\boldsymbol{\omega}_{uv}}^{(K)} < \frac{\beta}{4\alpha M}, \gamma_1 - \gamma_2 < p_{\boldsymbol{\omega}_{vu}}^{(K)} < \gamma_1 + \gamma_2$$

α, β are some constants γ_1, γ_2 are functions of $\left(M, p_{\boldsymbol{\omega}_{uv}}^{(K)}\right)$ (see Eqn. 11)

$$A\left(\Omega_M\left(p_{\boldsymbol{\omega}}^{(K)}, p_{\boldsymbol{\omega}_{uv}}^{\mathrm{BT}}\right)\right) = \frac{\beta^2}{6\alpha M^2}$$

Area of the region



Comparing B-T and P-L Models

Theorem 2 For all $M > 1$ and $K > 2$:
$$A\left(\Omega_M(p_{ij}^{BT}, p_{ik}^{BT})\right) > A\left(\Omega_M\left(p_{\boldsymbol{\omega}}^{(K)}, p_{\boldsymbol{\omega}_{uv}}^{\mathrm{BT}}\right)\right)$$

Takeaways

- Preference models with similar behaviors on the training set may assign significantly different probabilities to unseen preferences.
- Minor changes in the data distributions within the training set may lead to significant changes in the learned preference models.
- P-L models (with $K > 2$) are more robust than B-T model.
- Not just for value alignment, but wherever PM is used (e.g., Chatbot Arena)

Experiments

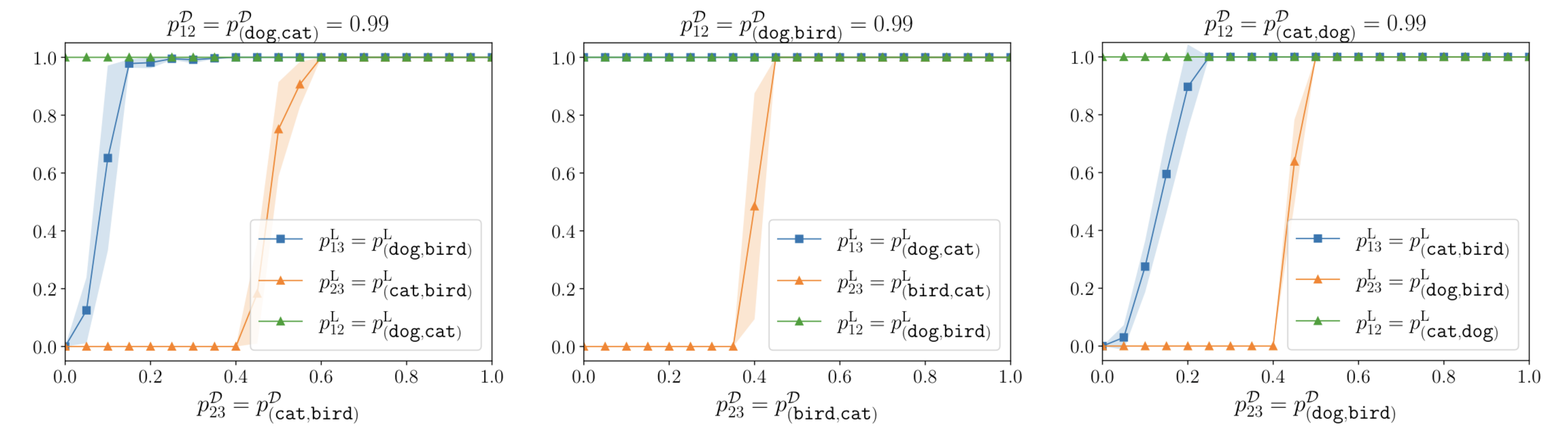
Strong Preferences are not Uncommon

Frequencies of preference probabilities assigned by reward models on `Anthropic/hh-rlhf`

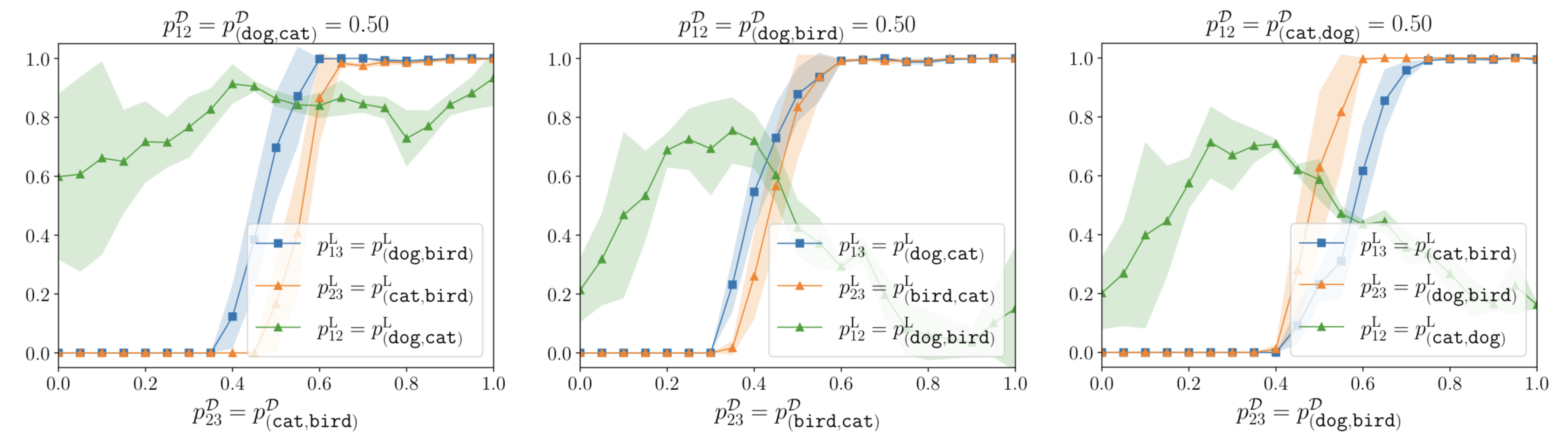
p_{wl}	Frequency of p_{wl}	
	`Llama-3.1-Nemotron-70B-Reward-HF`	`reward-model-deberta-v3-large-v2`
(0.00, 0.05)	1,184	22
[0.05, 0.10)	363	62
[0.10, 0.90)	3,636	7037
[0.90, 0.99)	1,574	1,264
[0.99, 1.00)	1,795	167
Total	8,552	

Sensitivities of Preference Models Manifest in Value Alignment

- Synthetic dataset: $\mathcal{O} = \{\text{dog}, \text{cat}, \text{bird}\}$ three preferences
- Set p_{12}^D to be 0.99 or 0.5 strong or moderate preferences
- Vary p_{23}^D from 0 to 1, resulting in changes in DPO-learned p_{23}^L
- Check the learned p_{13}^L, does it change proportionally to p_{23}^L?
- Studies LLMs: `Llama-3-8B-Instruct, zephyr-7b-alpha`



p_{12}^D =0.99: Despite small/no changes in p_{23}^L, a significant change in p_{13}^L occurs.



p_{12}^D =0.50: p_{13}^L tends to change proportionally to p_{23}^L.