TODO: Enhancing LLM Alignment with Ternary Preferences

Yuxiang Guo[1,2], Yin Lu[1,3], Bo Jiang[2], Jiaqi Zhang[1]
[1]Meituan Inc., [2]Beihang Univerisity, [3]University of Surrey

Code is available

Preference Data in Alignment

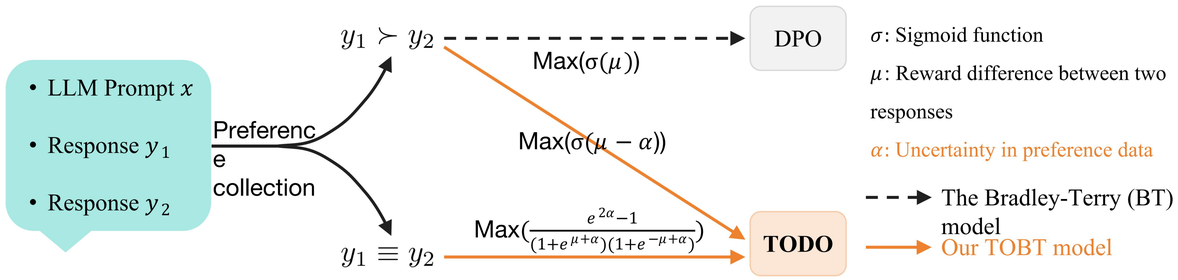
- **Inconsistent quality:** some pairwise preference data, often showing minimal discernible differences
- **Tie data:** tie data is not derived from additional annotations but is a natural occurrence
- **Limitation of BT model based alignment:** How to learn useful information from tie data and achieve nuanced preference modeling?

Tie-rank Oriented BT (TOBT) model

$$r_{12} = \frac{\lambda_1}{\lambda_1 + \phi\lambda_2} \qquad r_{(12)} = \frac{\lambda_1\lambda_2(\phi^2 - 1)}{(\lambda_1 + \phi\lambda_2)(\phi\lambda_1 + \lambda_2)}$$

- $\boldsymbol{\lambda}$: positive strength owned by responses
- $\phi = exp(\alpha)$: α encapsulates the inherent uncertainty and noise in the strength values λ_1 and λ_2

Tie-rank Oriented Direct Preference Optimization



1. Nuanced reward modeling: leveraging TOBT's ternary ranking system to improve preference alignment, offering a nuanced reward modeling

2. Learning more diversed information:

- When μ = 0, two responses obtain the same rewards, and DPO continues to update policy models towards the incorrect direction.

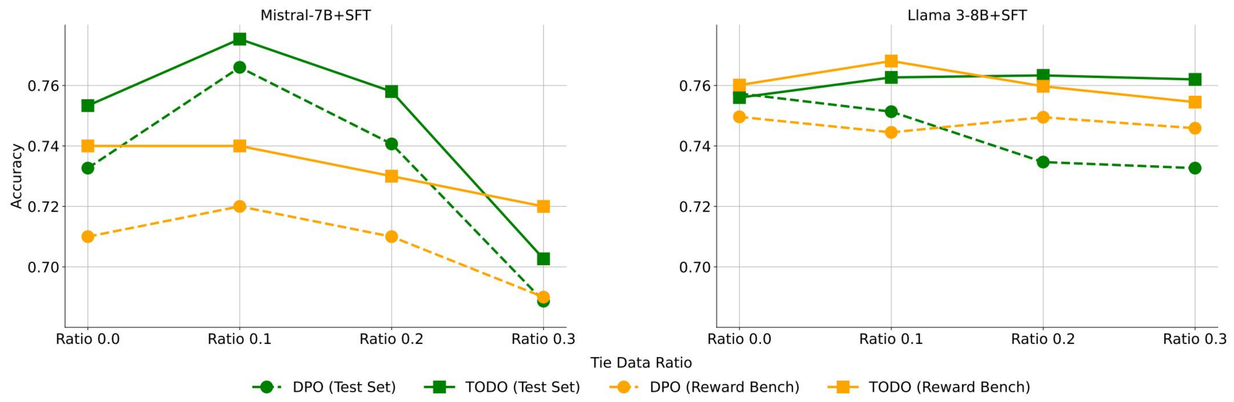
$$\nabla_\theta \mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_1,y_2)\sim\mathcal{D}}\Big[\underbrace{\beta\sigma(-\mu)}_{\text{always positive}}\Big[\nabla_\theta \log(\pi(y_1|x)) - \nabla_\theta \log(\pi(y_2|x))\Big]\Big].$$

- In contrast, TODO refrains from updating any parameters to maintain the consistent preference alignment of the tied responses.

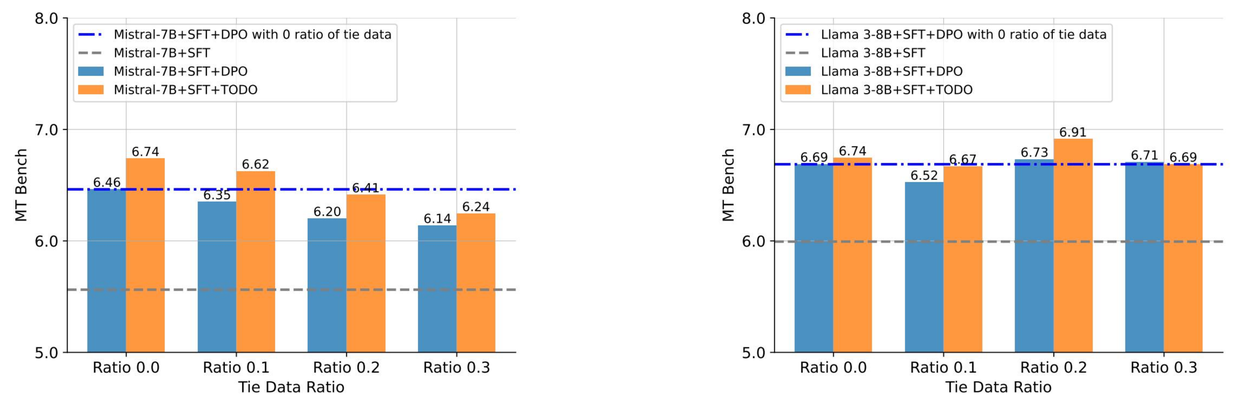
$$\nabla_\theta \mathcal{L}^t_{\text{TODO}}(\pi_\theta;\pi_{\text{ref}}) = -\mathbb{E}_{(x,y_1,y_2)\sim\mathcal{D}}\Big[\underbrace{G(\mu)\nabla_\theta \log(\pi(y_1|x))}_{\text{part 1}} + \underbrace{G(-\mu)\nabla_\theta \log(\pi(y_2|x))}_{\text{part 2}}\Big].$$

Experiments

Reward Modeing accuracy



MT Bench results



(a) Results on Mistral-7B (b) Results on Llama 3-8B

Results on reasoning-heavy tasks

Mistral-7B	Tie Data Ratio	Piqa	ARC-c	ARC-e	MMLU	Hellaswag	Winogrande	Average
DPO	0.0	77.15	49.87	69.56	59.44	73.55	67.88	66.24
TODO	0.0	**81.01**	**54.59**	**78.44**	**60.16**	**79.01**	**72.93**	**71.02**
DPO	0.1	79.87	53.82	77.21	60.07	79.04	73.48	70.58
TODO	0.1	**80.25**	**54.68**	**77.76**	**60.46**	**79.27**	**73.95**	**71.06**

Llama3-8B	Tie Data Ratio	Piqa	ARC-c	ARC-e	MMLU	Hellaswag	Winogrande	Average
DPO	0.0	79.27	54.59	76.58	63.25	79.05	71.74	70.75
TODO	0.0	**79.43**	**55.45**	**76.79**	**63.45**	**79.27**	**71.82**	**71.03**
DPO	0.3	79.05	55.54	76.11	63.34	78.17	70.88	70.52
TODO	0.3	**79.33**	**56.22**	**77.76**	**63.61**	**78.83**	**71.03**	**71.13**