Keep the Best, Forget the Rest: Reliable Alignment with Order-Aware Preference Optimization

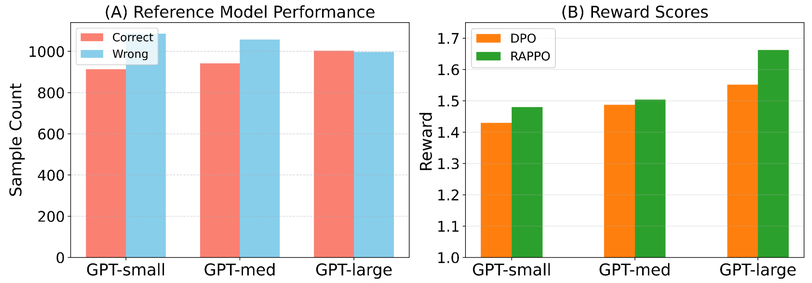
Jiahui Zhu[1], Yuanjie Shi[1], Xiyue Peng[2], Xin Liu[2], Yan Yan[1], Honghao Wei[1]

[1]Washington State University
[2]ShanghaiTech University

ICLR

Motivation

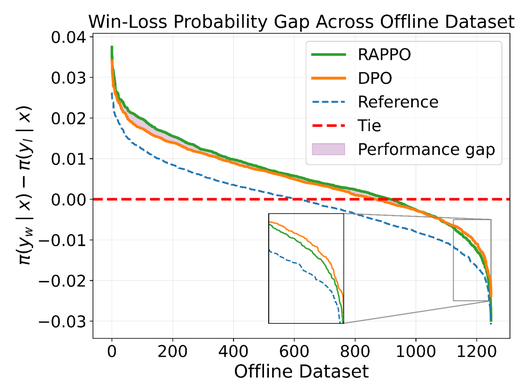
***Limitation**: Well-trained SFT models still misclassify preferences in pairwise data!!!!!!*



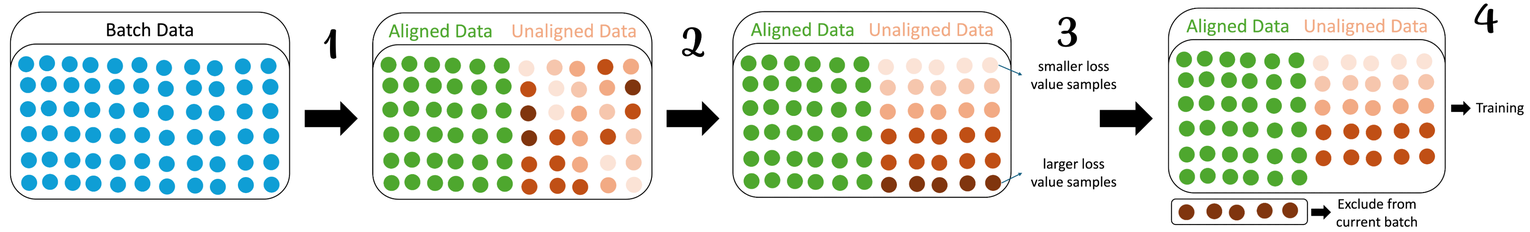
Can a simple, sample-aware modification to DPO mitigate reference-policy misalignment and thereby improve generalization?

Contribution

- **Reliable Filtering:** Dynamically discards ambiguous preference pairs.
- **Plug-and-Play:** Easily integrates into any DPO framework with minimal code.
- **Theoretical Guarantee**: Provably achieves a tighter learning bound than standard DPO.
- **Strong Performance**: Broadly outperforms baselines



Reliable Alignment for Preference Policy Optimization(RAPPO)



ALGORITHM RAPPO has following steps:

- **Sample:** Draw a mini-batch of preference data.
- **Partition:** Divide the batch into **Aligned** ($\frac{\pi_{ref}(y_w^i|x^i)}{\pi_{ref}(y_l^i|x^i)} > \tau$) and **Unaligned** ($\frac{\pi_{ref}(y_w^i|x^i)}{\pi_{ref}(y_l^i|x^i)} \leq \tau$) subsets.
- **Rank:** Order the samples within the **Unaligned** subset by their per-sample loss.
- **Filter & Update:** Discard the $\mathrm{Top\text{-}q}$ highest-loss **Unaligned** samples and train the model on the refined data.

THEOREM RAPPO has following advantages:

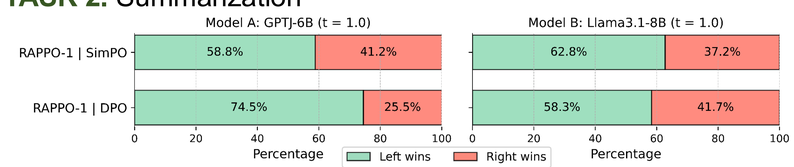
- **Maximal Descent.** expected first–order decrease for other sample selection ≤ expected first–order decrease for RAPPO
- **Minimal Variance.** variance of the signed decrease for other sample selection ≤ variance of the signed decrease for RAPPO
- **Uniform stability and generalization.** expected generalization error is bounded.

Experiments

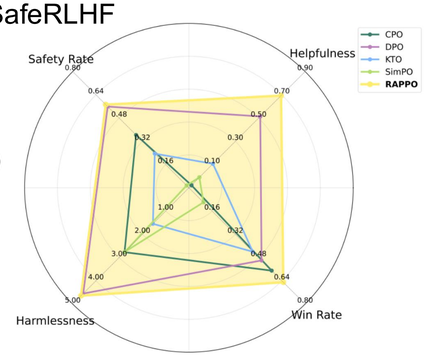
TASK 1. Controlled Generation: Sentiment & Toxicity

Algorithm	DPO	DPO-Offset	IPO	SimPO	RAPPO-1	RAPPO-2	RAPPO-3
Reward Score ↑	1.5513	1.5526	1.5446	1.5537	1.6625	1.6790	1.6811
Toxicity (%) ↓	6.30	8.11	6.49	7.49	2.64	2.60	2.28

TASK 2. Summarization



TASK 3. Large-Scale Evaluation on PKU-SafeRLHF



Sensitivity Analysis. # Removed Samples (q) & Selection Thresholds (τ)

Batch Size	q=1	q=2	q=4	q=8	Baseline (SimPO)
16 ($\tau = 1.0$)	1.7020	1.7481	1.7333	1.7111	1.6600
32 ($\tau = 0.8$)	1.6580	1.6720	1.6765	1.6510	1.5537
32 ($\tau = 1.0$)	1.6625	1.6790	1.6811	1.6432	1.5537
32 ($\tau = 1.2$)	1.6650	1.6805	1.6828	1.6595	1.5537

Resources

OpenReview

Source Code