Reward-Augmented Data Enhances Direct Preference Alignment of LLMs

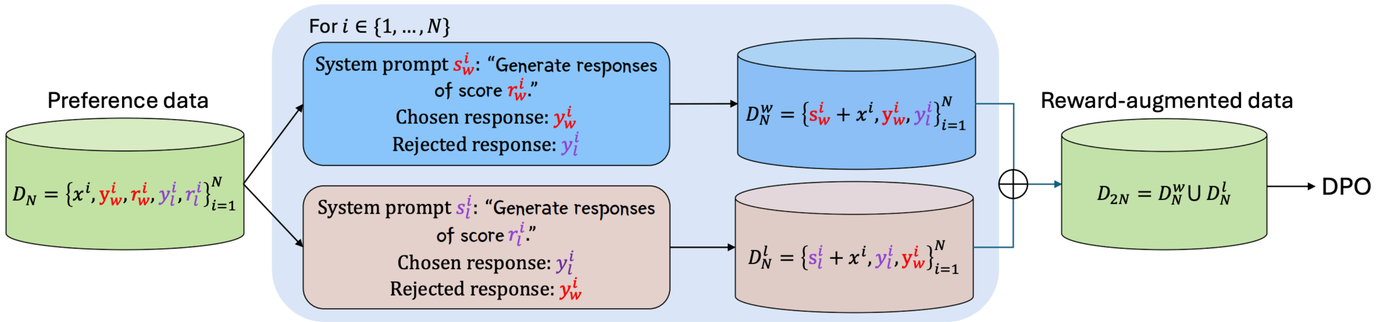
Shenao Zhang*[1], Zhihan Liu*[1], Boyi Liu[2], Yufeng Zhang[2],
Yingxiang Yang[2], Yongfei Liu[2], Liyu Chen[2], Tao Sun[2], Zhaoran Wang[1]

[1]Northwestern University, [2]ByteDance Seed

ICML International Conference On Machine Learning

arXiv

A simple data augmentation method for direct alignment!



Limitations of DPO

Relative preferences instead of response qualities

1. High-quality rejected text unnecessarily unlearned

response	y_1	y_2
$r(x,y)$	9	8
$\mathcal{D}_{N=1}$	$\{y_1 > y_2\}$	
$\pi^*(y \mid x)$	1	0

Optimal policy deterministically generates y_1

2. Low-quality chosen text indiscriminately learned

response	y_1	y_2	y_3
$r(x,y)$	9	1	0
$\mathcal{D}_{N=2}$	$\{y_1 > y_3\,,\, y_2 > y_3\}$		
$\pi^*(y \mid x)$	$1-a$	a	0

Optimal policy indiscriminately generates y_1 and y_2

3. Sparsity of optimal responses

Fail to characterize and generalize to these behaviors

Fix: Reward-Conditioned Policies

Learn from the full spectrum of qualities!

1. Learn from both high-quality responses

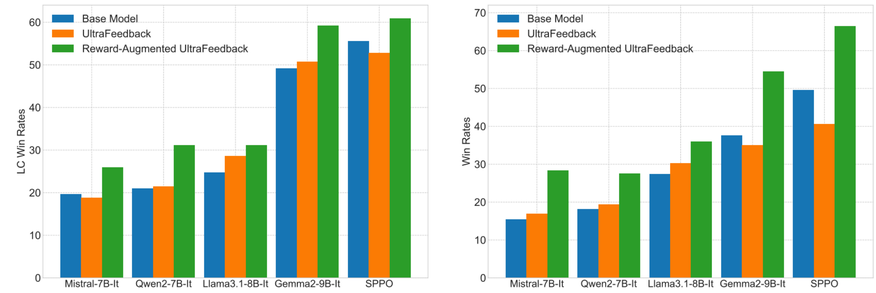
response	y_1	y_2
$r(x,y)$	9	8
$\mathcal{D}_{N=1}$	$\{y_1 > y_2\}$	
$\pi^*(y \mid x)$	1	0
$\pi^*(y \mid x, g=9)$	1	0
$\pi^*(y \mid x, g=8)$	0	1

2. Distinguish varying-quality responses

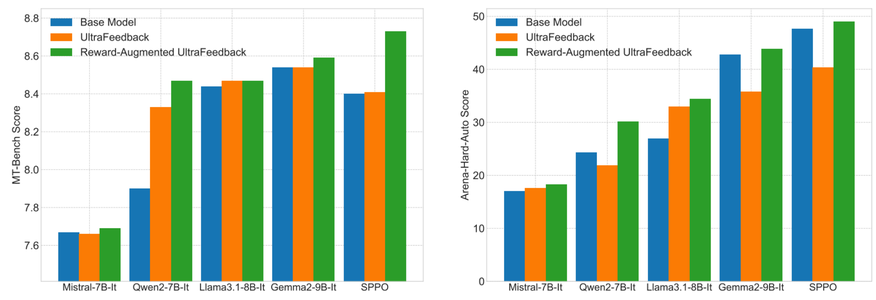
response	y_1	y_2	y_3
$r(x,y)$	9	1	0
$\mathcal{D}_{N=2}$	$\{y_1 > y_3\,,\, y_2 > y_3\}$		
$\pi^*(y \mid x)$	$1-a$	a	0
$\pi^*(y \mid x, g=9)$	1	0	0
$\pi^*(y \mid x, g=1)$	0	1	0
$\pi^*(y \mid x, g=0)$	0	0	1

3. Generalize with transferable features

Learning from g=8 and 9 helps generalize to g=10



(a) AlpacaEval 2.0 results. Left: Length-Controlled (LC) win rates. Right: Win rates.



(b) MT-Bench average score. (c) Arena-Hard-Auto score.

Reward augmentation gets more juice out of the data.

1. SPPO + DPO ↓ but SPPO + DPO (RA) ↑

	LC WR	WR	MT	Arena
SPPO	55.60	49.61	8.40	47.6
+DPO (UF)	52.75	40.58	8.41	40.4
+DPO (RA)	**60.97**	**66.41**	**8.73**	**49.0**

2. Another DPO round on implicit-reward-relabeled data enhances the performance

	LC WR	WR	MT	Arena
Qwen2-7B-It	20.93	18.22	7.90	24.3
+DPO (UF)	21.46	19.35	8.33	21.9
+DPO (RA)	31.17	27.58	8.47	**30.1**
+DPO (IRA)	**32.61**	**29.15**	**8.49**	28.3

…Find more ablations in our paper!