Bridging and Modeling Correlations in Pairwise Data for Direct Preference Optimization

Yuxin Jiang[1,2], Bo Huang[1,2], Yufei Wang[3], Xingshan Zeng[3], Liangyou Li[3], Yasheng Wang[3], Xin Jiang[3], Lifeng Shang[3], Ruiming Tang[3], Wei Wang[1,2]

1. HKUST(GZ) 2. HKUST 3. Huawei Noah's Ark Lab

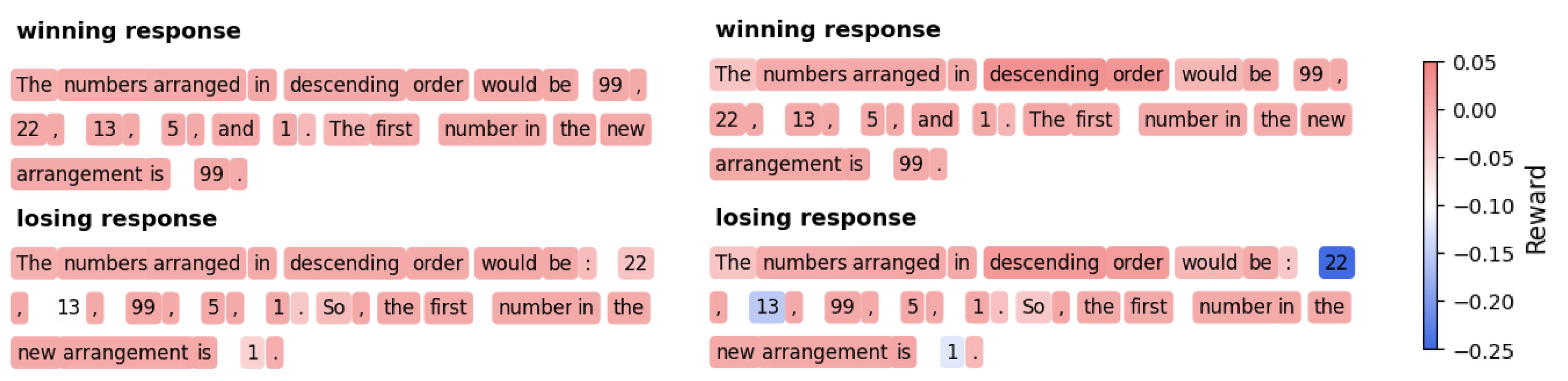
Paper Code

Introduction

- In DPO, the generation of y_w and y_l are typically produced without mutual visibility, resulting in a lack of strong correlation or relevance between them.

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

- Thus, the model may struggle to identify nuanced yet significant distinctions that differentiate superior responses from inferior ones, leading to suboptimal alignment performance.



(a) DPO (b) DPO-BMC

Methodology

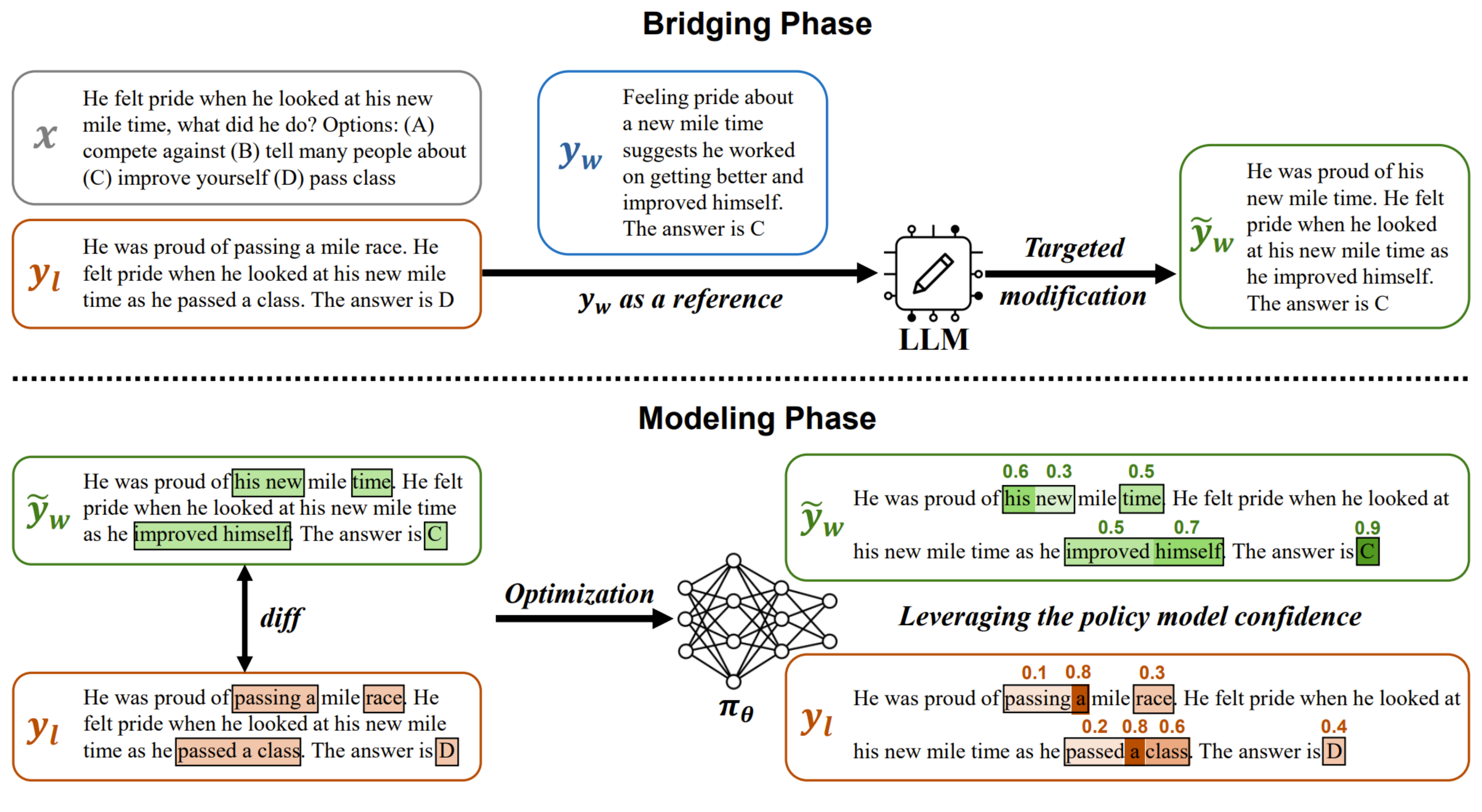


Figure 1: Overview of our proposed BMC framework. (1) In the Bridging Phase, we utilize an off-the-shelf LLM to make *targeted modifications* of losing response y_l on undesired tokens, with the winning response y_w serving as a reference. Therefore, the synthesized pseudo-winning response $\tilde{y}_w$ is highly correlated with y_l. (2) In the Modeling Phase, we model the correlations between $\tilde{y}_w$ and y_l by *dynamically* emphasizing the rewards of their varied tokens ($diff(\tilde{y}_w \mid y_l)$ and $diff(y_l \mid \tilde{y}_w)$), leveraging the policy model confidence (numbers indicated above tokens) during training.

- **Bridging Phase:**

$\mathrm{LLM}(I, x, y_w, y_l) \rightarrow \tilde{y}_w$, where I is the instruction of targeted modification.

- **Modeling Phase:**

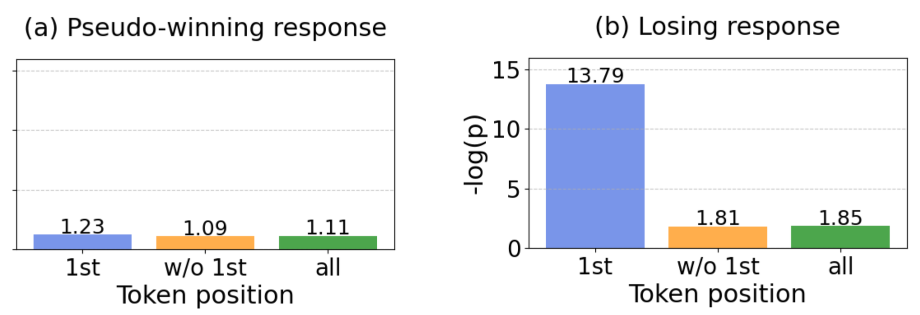
$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(\tau_w,\tau_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\sum_{t=0}^{N-1}\log\frac{\pi_\theta(a_w^t \mid s_w^t)}{\pi_{\text{ref}}(a_w^t \mid s_w^t)} - \beta\sum_{t=0}^{M-1}\log\frac{\pi_\theta(a_l^t \mid s_l^t)}{\pi_{\text{ref}}(a_l^t \mid s_l^t)}\right)\right] \text{(token-level MDP format)}$$

leveraging the policy model confidence

$$\mathcal{L}_{\text{DPO-BMC}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,\tilde{y}_w,y_l)\sim\tilde{\mathcal{D}}}\left[\log\sigma\left(\beta\sum_{\tilde{y}_w^t\in\tilde{y}_w}\lambda_{\tilde{y}_w^t}\log\frac{\pi_\theta(\tilde{y}_w^t \mid \tilde{y}_w^{<t}, x)}{\pi_{\text{ref}}(\tilde{y}_w^t \mid \tilde{y}_w^{<t}, x)} - \beta\sum_{y_l^t\in y_l}\lambda_{y_l^t}\log\frac{\pi_\theta(y_l^t \mid y_l^{<t}, x)}{\pi_{\text{ref}}(y_l^t \mid y_l^{<t}, x)}\right)\right],$$

$$\lambda_{\tilde{y}_w^t} = \begin{cases} 1 + \min\left(sg\left(\frac{1}{\pi_\theta(\tilde{y}_w^t \mid \tilde{y}_w^{<t}, x)}\right), \delta\right), & \text{if } \tilde{y}_w^t \in diff(\tilde{y}_w \mid y_l) \\ 1, & \text{otherwise} \end{cases}$$

$$\lambda_{y_l^t} = \begin{cases} 1 + \min\left(sg\left(\frac{1}{\pi_\theta(y_l^t \mid y_l^{<t}, x)}\right), \delta\right), & \text{if } y_l^t \in diff(y_l \mid \tilde{y}_w) \\ 1, & \text{otherwise} \end{cases}$$



Experiments

Method	Question-Answering Tasks					Mathematical Reasoning Tasks				
	ECQA	QASC	OBQA	StrategyQA	Avg.	GSM8k	MATH	MAWPS	TabMWP	Avg.
SFT	72.8	54.5	51.8	56.9	59.0	55.8	11.6	80.3	42.8	47.6
FIGA	70.3	52.5	51.7	48.6	55.8	54.1	9.8	75.5	39.0	44.6
IPO	71.5	58.9	53.6	58.4	60.6	57.2	12.1	82.2	42.5	48.5
OPRO	69.8	55.1	51.4	57.2	58.4	56.0	12.4	80.8	41.3	47.6
R-DPO	73.5	59.5	55.4	58.8	61.8	56.9	12.0	81.9	42.2	48.2
SimPO	71.9	56.7	52.2	55.4	59.1	57.5	12.7	81.8	<u>43.5</u>	48.9
DPO	73.1	58.8	55.6	57.8	61.3	56.3	12.3	81.2	43.4	48.3
DPO (CW)	72.5	58.6	55.2	57.3	60.9	55.9	11.8	80.7	42.8	47.8
DPO (EW)	72.9	59.4	55.8	57.9	61.5	56.5	12.0	80.9	43.4	48.2
DPO-BMC	**75.9**	**63.0**	**60.4**	**61.0**	**65.1**	**58.4**	**13.0**	**83.1**	**43.8**	**49.6**
DPO-BC	<u>75.7</u>	<u>62.0</u>	56.0	<u>60.1</u>	<u>63.4</u>	<u>57.6</u>	<u>12.7</u>	<u>82.8</u>	43.4	<u>49.1</u>
DPO-MC	74.8	60.0	<u>56.4</u>	58.8	62.5	57.2	12.5	82.4	43.0	48.8

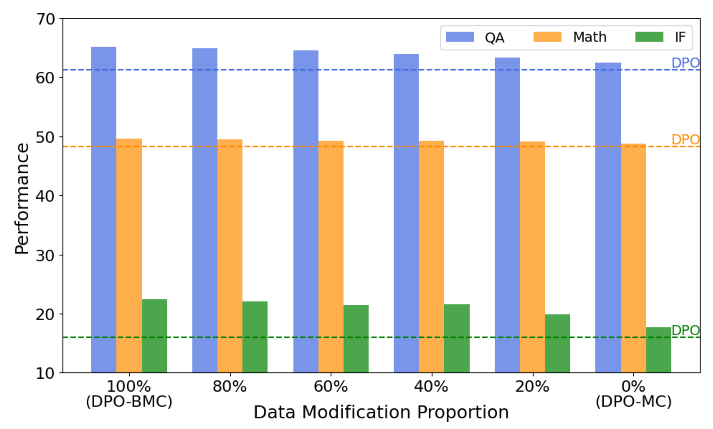
Method	Llama3-8B-Base					Mistral-7B-Base				
	AlpacaEval 2			Arena-Hard		AlpacaEval 2			Arena-Hard	
	LC (%)	WR (%)	Avg. len	WR (%)	Avg. len	LC (%)	WR (%)	Avg. len	WR (%)	Avg. len
SFT	7.5	4.7	956	2.6	414	8.1	5.9	998	2.2	454
FIGA	8.4	4.2	1,199	5.1	416	7.0	4.9	1,378	2.5	461
IPO	13.4	9.8	1,430	14.0	477	12.5	10.8	1,588	8.5	522
ORPO	12.5	11.4	1,793	11.7	573	14.5	11.5	1,630	9.4	566
R-DPO	17.1	14.4	1,801	17.6	582	16.0	12.3	1,521	10.4	529
SimPO	<u>21.3</u>	**18.9**	1,718	**26.6**	562	16.8	<u>14.4</u>	1,906	**18.4**	615
DPO	16.0	14.8	1,713	17.6	559	15.1	13.3	1,657	13.6	540
DPO (CW)	15.2	14.0	1,756	17.1	570	14.5	12.9	1,647	13.0	532
DPO (EW)	17.2	15.6	1,702	18.2	566	15.3	13.4	1,668	13.9	549
DPO-BMC	**22.4**	<u>16.8</u>	1,285	<u>18.1</u>	406	**20.8**	**16.6**	1,317	<u>17.6</u>	488
DPO-BC	20.6	14.4	1,269	16.8	422	<u>18.6</u>	13.8	1,489	15.9	502
DPO-MC	17.7	15.2	1,890	17.9	579	16.4	14.3	1,712	15.4	551

Ablation Study

- **Data synthesis methods**

Data Synthesis	Training Data	QA	Math	IF
$y_l \xrightarrow{y_w} \tilde{y}_w$ (ours)	$(\tilde{y}_w, y_l)$	**65.1**	**49.6**	**22.4**
$y_l \longrightarrow \tilde{y}_w$	$(\tilde{y}_w, y_l)$	64.3	49.2	19.8
$y_w \xrightarrow{y_l} \tilde{y}_l$	$(y_w, \tilde{y}_l)$	64.6	48.7	18.9
$y_w \longrightarrow \tilde{y}_l$	$(y_w, \tilde{y}_l)$	63.9	48.6	17.6

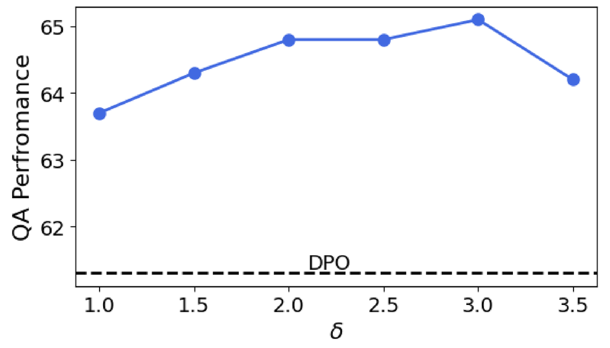
- **Data modification proportion**



- **Data synthesis LLMs**

Method	LLM for Targeted Modification	QA	Math	IF
SFT	–	56.9	47.6	7.5
DPO	–	61.3	48.3	16.0
DPO-BMC	Llama3-70B-Instruct	64.6	49.4	21.8
DPO-BMC	gpt-4-0125-preview	**65.1**	**49.6**	**22.4**

- **δ in the Modeling Phase**



Why BMC Works?

(y_w, y_l)	Edit Distance	LC (%)	Grad Norm
split 1	0.57	7.68	3.31
split 2	0.70	9.49	4.85
split 3	0.73	10.50	4.86
split 4	0.76	10.01	5.33
split 5	0.83	8.57	6.31
split 6	0.95	7.91	13.00

(a) DPO

(y_w, y_l)	Edit Distance	LC (%)	Grad Norm
split 1	0.57	9.40	5.70
split 2	0.70	12.49	8.39
split 3	0.73	13.27	8.66
split 4	0.76	11.47	9.03
split 5	0.83	9.81	8.44
split 6	0.95	9.90	9.04

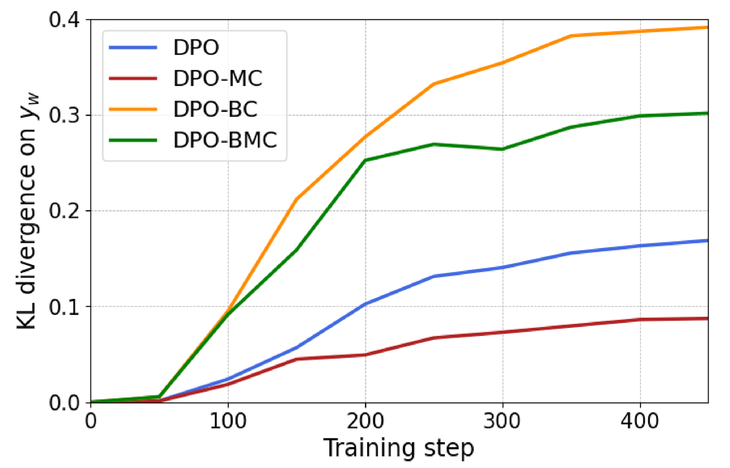
(b) DPO-MC

$(\tilde{y}_w, y_l)$	Edit Distance	LC (%)	Grad Norm
split 1	0.45	10.82	3.47
split 2	0.52	10.87	4.80
split 3	0.56	12.54	5.20
split 4	0.61	14.34	5.39
split 5	0.70	13.24	6.98
split 6	0.84	10.59	9.67

(c) DPO-BC

$(\tilde{y}_w, y_l)$	Edit Distance	LC (%)	Grad Norm
split 1	0.45	11.21	5.26
split 2	0.52	11.49	7.33
split 3	0.56	11.47	7.70
split 4	0.61	14.38	8.17
split 5	0.70	15.28	7.65
split 6	0.84	12.29	8.75

(d) DPO-BMC



- Bridging Phase fosters tailored learning toward critical differences in preference data.
- Modeling Phase promotes a balanced optimization landscape by encouraging challenging distinctions while reinforcing learned patterns.