Disentangling Length Bias in Preference Learning via Response-Conditioned Modeling

Jianfeng Cai, Jinhua Zhu, Ruopei Sun, Yue Wang, Li Li, Wengang Zhou, Houqiang Li**

* corresponding authors

University of Science and Technology of China

Zhongguancun Academy

ICLR

Abstract & Contributions

- We identify a central failure mode in RLHF preference modeling: reward models can exploit spurious correlates, with length bias being a prominent example, and aligned LLMs often struggle to follow explicit length control instructions.
- We propose a response-conditioned preference modeling framework (**Rc-BT**) that explicitly conditions comparisons on response attributes to decouple true quality preference from length-related shortcuts, improving both length-bias robustness and length controllability.
- Building on Rc-BT, we develop **Rc-RM** and **Rc-DPO** for reward modeling and preference optimization, and empirically show consistent gains (reduced length bias and better length instruction following) across multiple models and datasets.

Motivation: Two Key Challenges

Challenge 1 — Length Bias in Reward Models
Standard BT reward models implicitly correlate reward score with response length, assigning higher scores to longer replies even when quality is lower. This misdirects downstream policy optimization.

Challenge 2 — Poor Length Instruction Following
Fine-tuned LLMs consistently fail to comply with explicit length constraints, despite showing implicit sensitivity to length cues during pretraining and RLHF.

Insight: LLMs are sensitive to length, we leverage this to disentangle length from semantics!

Core Formulation

Standard Bradley-Terry (BT):

$$p^*(y_c \succ y_r|x) = \frac{\exp\left(r^*(x, y_c)\right)}{\exp\left(r^*(x, y_c)\right) + \exp\left(r^*(x, y_r)\right)}$$

Response-Conditioned BT (Rc-BT):

$$p^*(x_c \succ x_r|y) = \frac{\exp\left(r^*(x_c, y)\right)}{\exp\left(r^*(x_c, y)\right) + \exp\left(r^*(x_r, y)\right)}$$

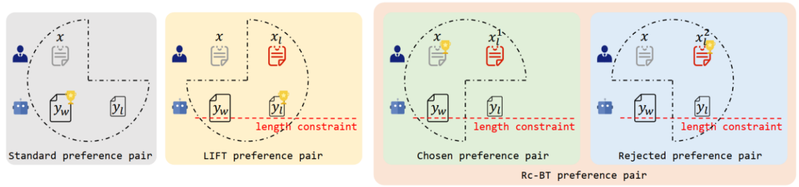
Rc-BT Instantiation for Length:

$$p^*(x \succ x_l^1|y_c) = \frac{\exp\left(r^*(x, y_c)\right)}{\exp\left(r^*(x, y_c)\right) + \exp\left(r^*\left(x_l^1, y_c\right)\right)}$$

$$p^*(x_l^2 \succ x|y_r) = \frac{\exp\left(r^*\left(x_l^2, y_r\right)\right)}{\exp\left(r^*\left(x_l^2, y_r\right)\right) + \exp\left(r^*(x, y_r)\right)}$$

where $x_l^1, x_l^2 = x + length\ constraint$ — **explicitly conditioned on length requirement**

Rc-BT Framework



End-to-End Training Pipeline

Original Dataset D_{rm} (x, y_c, y_r) → augment → **Augmented Dataset D_{Rc}** $(x, x_l^1, y_c) \cup (x_l^2, x, y_r)$

Training on $D_{rm} \cup D_{Rc}$

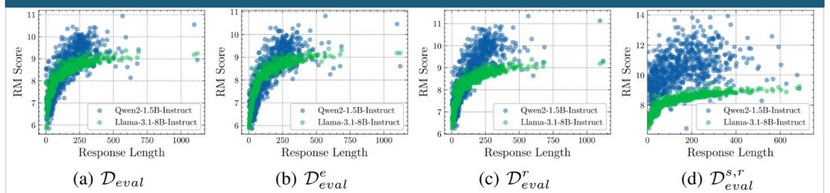
Rc-BT Model $p^*(x_c \succ x_r|y)$

Rc-RM
Reward Modeling

Rc-DPO
Policy Optimization

- Plug-and-play into RM and DPO stages
- No external LLM required for augmentation
- Generalizes across base models & datasets

Preliminary Explorations I: Length Bias Indeed Exists



(a) $\mathcal{D}_{eval}$ (b) $\mathcal{D}_{eval}^{e}$ (c) $\mathcal{D}_{eval}^{r}$ (d) $\mathcal{D}_{eval}^{s,r}$

- Reward models consistently assign higher scores to longer responses across different datasets and models, confirming that length bias is a systematic and pervasive issue in preference learning.

Preliminary Explorations II: Length Instructions Are Easily Learned

Model	Variant	Quality Eval Acc (%)	Length Eval Acc (%)
Qwen2-1.5B-Base	LIFT-plus$_2^{reverse}$	58.78	85.58
	LIFT-plus$_2^{noreverse}$	59.41	87.78
	LIFT-plus$_2^{empty}$	58.33	89.71
Qwen2-1.5B-Instruct	LIFT-plus$_2^{reverse}$	60.11	86.54
	LIFT-plus$_2^{noreverse}$	63.17	85.21
	LIFT-plus$_2^{empty}$	61.02	85.53
Llama-3.1-8B-Base	LIFT-plus$_2^{reverse}$	51.34	91.32
	LIFT-plus$_2^{noreverse}$	57.71	88.14
	LIFT-plus$_2^{empty}$	50.27	88.78
Llama-3.1-8B-Instruct	LIFT-plus$_2^{reverse}$	60.11	95.19
	LIFT-plus$_2^{noreverse}$	56.65	92.31
	LIFT-plus$_2^{empty}$	57.97	90.71

- Models can quickly learn explicit length instructions, but this often leads to **overfitting** to length at the expense of semantic quality **using standard BT framework**.

Experiments: Reward Model And DPO Policy Optimization

Model	Baseline	ODIN	R-DA	Rc-RM
Qwen2-1.5B-Base	59.14	56.12	60.17	**69.55**
Qwen2-1.5B-Instruct	60.75	63.56	61.27	**71.47**
Qwen2.5-7B-Base	54.26	60.90	63.46	**70.74**
Qwen2.5-7B-Instruct	59.31	67.55	66.34	**73.07**
Llama-3.1-8B-Base	59.04	60.37	65.17	**70.51**
Llama-3.1-8B-Instruct	55.59	60.90	60.78	**72.44**
Gemma-2-9B-it	53.45	55.85	55.21	**63.56**
Qwen2.5-14B-Instruct	65.57	76.22	75.14	**81.70**

Reward Model:
Our Rc-RM consistently achieves the **best** reward-model performance across model families and scales, outperforming both baseline and debiasing baselines (e.g., ODIN, R-DA), indicating a more faithful preference signal that is less driven by length.

DPO Policy Optimization:
Our Rc-DPO delivers the **highest** quality win ratios, while producing shorter responses than other baselines and provides the best quality and length trade-off, which shows that policy improvements come from better semantics rather than exploiting longer outputs (i.e., effective length-bias mitigation).

Metrics	Baseline	LIFT-plus	R-DPO	Dr.DPO	SimPO	CPO	Rc-DPO
Qwen2.5-7B-Base							
Quality Win Ratio (%)	33.54	31.67	40.40	39.74	41.26	40.07	**45.39**
Response Length	517.30	184.17	583.18	311.49	286.54	254.86	208.42
Qwen2.5-7B-Instruct							
Quality Win Ratio (%)	28.43	25.69	34.16	36.12	37.16	35.72	**44.63**
Response Length	261.32	195.80	235.39	247.19	281.26	257.95	228.24
Llama-3.1-8B-Base							
Quality Win Ratio (%)	46.30	40.15	49.13	51.12	54.38	50.86	**58.10**
Response Length	435.98	157.80	465.72	347.29	309.86	288.64	202.01
Llama-3.1-8B-Instruct							
Quality Win Ratio (%)	42.52	47.88	42.64	52.18	58.13	54.67	**64.34**
Response Length	247.74	153.77	215.82	229.17	218.46	244.04	204.77