Mitigating Length Bias in RLHF Through a Causal Lens

Hyeonji Kim[1] Sujeong Oh[2] Sanghack Lee[1*] ([1]Seoul National University)

Association for the Advancement of Artificial Intelligence

Seoul National University
Graduate School of Data Science

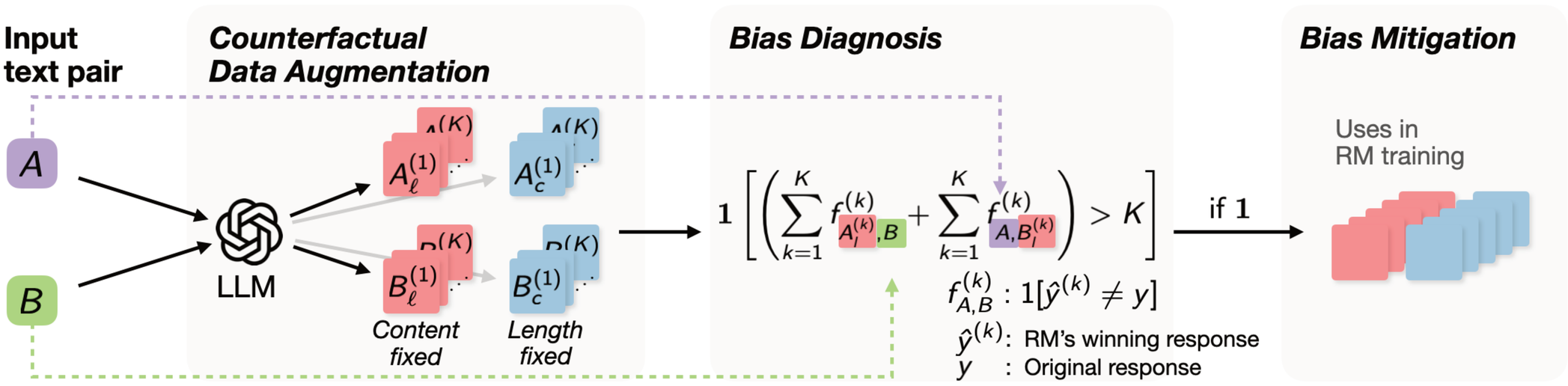

Figure 1: Method overview.

Causal Interpretation of Length Bias

Length Bias: Reward models favor longer responses even with worse information.

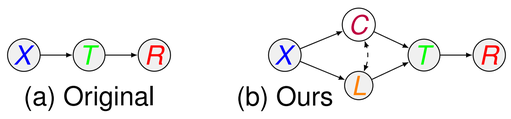

Figure 2: Causal view of length bias.

(a) **Common perspective**: prompt $X \rightarrow$ response $T \rightarrow$ reward R.

(b) **Causal perspective**: response T depends on content C and length L, potentially interacting.

Motivation for Counterfactuals

Problem. Semantic content C and response length L co-vary $\rightarrow$ observational comparisons cannot isolate the **causal effect of length on reward**.

Key Idea: Counterfactuals. Generate synthetic response pairs to answer questions that natural data cannot:

"What would the reward have been if the response length were different?"

Our Approach. Counterfactual data are constructed by independently intervening on content C and length L. Training on **equal-length counterfactual pairs** forces the reward model to ground preferences in **semantic content**, not verbosity.

Assumptions.

- Each response T decomposes into latent content C and observable length L, which affect reward only through the response.
- It is possible to modify one factor (e.g., L) while approximately preserving the other (e.g., C).
- Given two responses of similar length, preference differences reflect semantic content quality.

Length Bias Mitigation Pipeline

Goal. Mitigate length bias by training reward models on counterfactual comparisons that causally decouple content from length.

Three-Stage Framework

- **Counterfactual Augmentation**: Length/content intervention with the other factor fixed.
- **Bias Diagnosis**: Length-driven preference flips under content-fixed interventions.
- **Bias Mitigation**: Counterfactual supervision that neutralizes verbosity.

1. **Counterfactual Data Augmentation**: Content/length intervention $\rightarrow$ comparisons at fixed length.
   - **Length-Fixed**: Vary semantic content / align length.
   - **Content-Fixed**: Vary verbosity / preserve semantic content.

   Verification: Length via token bins; content via semantic filtering.
2. **Bias Diagnosis**:
   - **Rule-Based**: Longer wins with same or worse content
   - **Key signal**: Preference flips under content-preserving length changes
   - **Flip Ratio**: Majority flips ($F > 0.5$) indicate length bias
3. **Mitigating Length Bias**:
   - **Flipped Supervision**: Content-fixed counterfactuals $\Rightarrow$ length neutralized
   - **Semantic Grounding**: Length-fixed variants $\Rightarrow$ supervise semantic differences
   - **Training**: Flipped content-fixed + Length-fixed augmented dataset
   - **Effect**: Content-driven preference learning

Takeaway. Counterfactual supervision that **decouples content from length** reveals and mitigates length bias, yielding **length-invariant reward learning** in RLHF.

Experiments

Reward Model Evaluation

Model	RewardBench-1					RewardBench-2						Chatbot Arena
	Chat	Chat Hard	Safety	Reasoning	Avg	Factuality	PIF	Math	Safety	Focus	Avg	LC Accuracy
HRO	0.718	0.485	0.334	0.420	0.486	0.364	**0.275**	**0.350**	0.240	**0.238**	0.250	0.249
ODIN	0.499	0.487	0.514	0.485	0.496	0.301	0.263	0.230	0.154	0.147	0.219	0.463
CDA_OpenLM*	0.466	0.493	0.504	0.482	0.486	0.416	0.257	0.311	0.270	0.133	0.278	**0.508**
CDA_LoRA*	**0.732**	0.496	0.332	0.427	0.497	0.361	0.244	0.336	0.267	0.232	**0.288**	0.248
CDA_HRO*	0.491	**0.510**	**0.529**	**0.495**	**0.506**	**0.461**	0.197	0.244	**0.281**	0.199	0.276	0.493

Table 1: Accuracy on **RewardBench-1**, **RewardBench-2**, and length-controlled (LC) accuracy from **Chatbot Arena**.

Key Findings on Reward Models

- **RewardBench:** CDA achieves the highest average accuracy and balanced performance.
- **Length-Controlled:** CDA retains accuracy when length cues are removed.
- **Reward–Length:** Baselines correlate reward with response length; CDA is length-invariant.

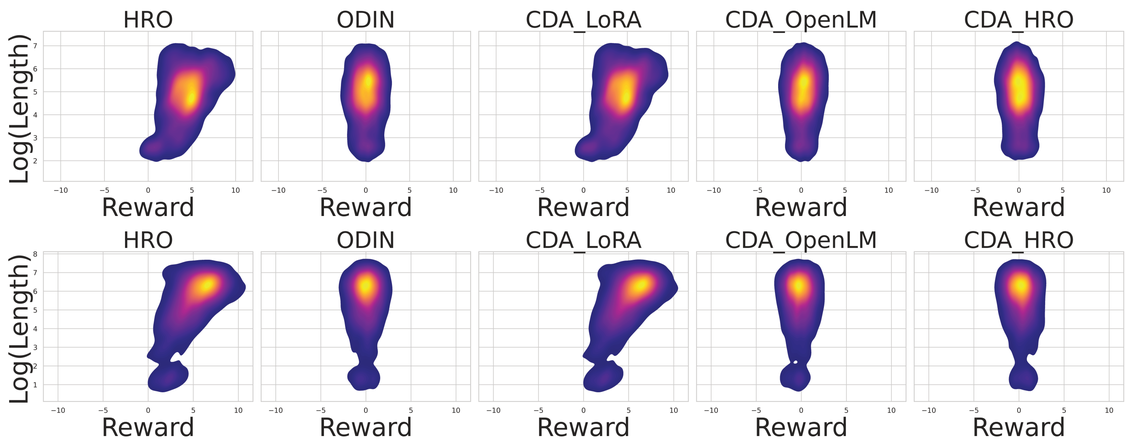

Figure 3: Reward distribution on RewardBench-1 (top) and RewardBench-2 (bottom).

Policy Evaluation with PPO

Metric	OpenLM	SFT	PPO_HRO	ODIN	PPO_CDA_OpenLM*	PPO_CDA_HRO*
LC Winrate	8.47	16.97	18.97	12.19	36.06	**37.18**
Winrate	9.94	25.71	28.45	11.34	30.69	**32.55**
Avg. length	1385	2061	2048	1026	1072	1118

Table 2: AlpacaEval Winrates.

Key Finding on PPO Policies trained with CDA reward models achieve higher length-controlled winrates while generating shorter responses.

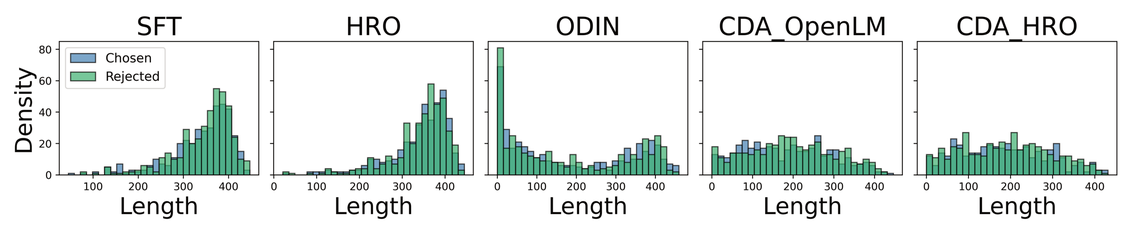

Figure 4: Length distribution of model outputs.

Conclusion

1. Counterfactual interventions on content and length disentangle semantic quality from verbosity in RLHF reward models.
2. Training on these counterfactuals mitigates length bias and yields more concise, informative, and robust preferences.

The 40th Annual AAAI Conference on Artificial Intelligence

Acknowledgements: This work was supported by NRF (RS-2023-00211904/50%, RS-2023-00222663/50%) grant funded by the Korean government.