Mitigating Length Bias in RLHF Through a Causal Lens

Hyeonji Kim[1], **Sujeong Oh**[1], **Sanghack Lee**[1*]

[1]Graduate School of Data Science, Seoul National University

Abstract

Reinforcement learning from human feedback (RLHF) is widely used to align large language models (LLMs) with human preferences. However, RLHF-trained reward models often exhibit length bias—a systematic tendency to favor longer responses by conflating verbosity with quality. We propose a causal framework for analyzing and mitigating length bias in RLHF reward modeling. Central to our approach is a counterfactual data augmentation method that generates response pairs designed to isolate content quality from verbosity. These counterfactual examples are then used to train the reward model, enabling it to assess responses based on content quality independently of verbosity. Specifically, we construct (1) length-divergent pairs with similar content and (2) content-divergent pairs of similar length. Empirical evaluations show that our method reduces length bias in reward assignment and leads to more concise, content-focused outputs from the policy model. These findings demonstrate that the proposed approach effectively reduces length bias and improves the robustness and content sensitivity of reward modeling in RLHF pipelines.

1 Introduction

Large language models (LLMs) have demonstrated remarkable performance across a wide range of natural language tasks (Brown et al. 2020; Liang et al. 2023; Chowdhery et al. 2023). Reinforcement learning from human feedback (RLHF) (Ziegler et al. 2019; Stiennon et al. 2020) has become the dominant approach for aligning LLM behavior with human preferences (Ouyang et al. 2022). Despite its success, RLHF often inherits and amplifies systematic biases inherently present in human preference data, with *length bias* being one of the most persistent issues (Ouyang et al. 2022; Shen et al. 2023; Saito et al. 2023). Length bias refers to the tendency of reward models to assign higher scores to longer responses, even when informativeness and relevance are comparable or worse. This bias can significantly distort model behavior and user experience (Singhal et al. 2024).

Recent studies have empirically demonstrated that both models and human annotators are susceptible to verbosity bias (Saito et al. 2023; Shen et al. 2023), often preferring longer responses even when content is held constant. When such preferences are implicitly encoded into reward models, RLHF-trained LLMs tend to prioritize verbosity over clarity, often resulting in unnecessarily long and less effective outputs that may degrade user experience. This phenomenon may arise because reward models leverage spurious correlations in data that fail to capture the true quality of the output (Stiennon et al. 2020; Singhal et al. 2024; Huang et al. 2025).

Several methods have been proposed to mitigate length bias. For instance, Chen et al. (2024); Wang et al. (2025) operate on learned representations by regularizing the reward model, and Liu et al. (2025); Cai et al. (2025) generate randomized or loosely controlled response pairs to reduce sensitivity to length. However, these approaches often lack the ability to explicitly disentangle verbosity from semantic quality, still leaving the reward model vulnerable to spurious correlations between response length and the reward.

To guide the reward model toward learning preferences based on content quality rather than surface length features, we adopt a causal perspective on length bias. Without such a framework, it is difficult to separate genuine effects from misleading patterns. For example, although there is a strong correlation between a country's per capita chocolate consumption and its number of Nobel laureates, both are influenced by a third factor such as national wealth. This classic example highlights the risk of relying solely on observational correlations without causal reasoning.

To address this, we propose a *counterfactual data augmentation* framework that enables reward models to disentangle content quality from response length. While counterfactual data augmentation has been used to mitigate spurious correlations in classification tasks (Kaushik, Hovy, and Lipton 2020), our work extends this causal idea to the RLHF reward-modeling by asking: "How would the reward change if the same content were expressed more concisely?" To answer this, we generate two types of counterfactual preference pairs: (1) semantically equivalent responses of different lengths, and (2) semantically different responses of similar lengths. These comparisons isolate the effects of content and verbosity, guiding the reward model to develop preferences based on semantic quality rather than length. Our contributions are:

- We identify a key limitation of existing approaches to mitigating length bias: their limited ability to disentangle verbosity from semantic quality due to reliance on spurious correlation. Our analysis shows that, without explicit

*Corresponding author.

interventions, current methods often conflate response length with informativeness.

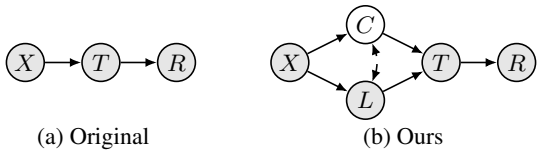
- We propose a novel counterfactual data augmentation framework which enables reward models to generate rewards by separating content quality from response length. Training on carefully constructed response pairs that isolate one factor at a time, our method facilitates content-based preference learning during reward model training.
- We empirically demonstrate that our approach effectively mitigates length bias and leads to more robust content-sensitive reward modeling in LLM, as evidenced by comprehensive evaluations across multiple benchmarks.

2 Preliminaries

Length Bias in Reward Model Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022) aligns language models with human preferences via fine-tuning on pairwise comparisons. Common approaches include Proximal Policy Optimization (PPO) (Schulman et al. 2017), which uses a learned reward model, and Direct Preference Optimization (DPO) (Rafailov et al. 2023), which directly optimizes preferences without an explicit reward model. See Appendix A[1] for further RLHF details.

Several methods have been proposed to mitigate length bias in reward modeling, primarily through architectural or training-time interventions. ODIN (Chen et al. 2024) uses a dual-head reward model to isolate semantic and stylistic features, while RRM (Liu et al. 2025) augments preference data with length perturbations to promote robustness. Although both reduce surface-level sensitivity to response length, they do *not* perform controlled interventions on length itself, and may suppress stylistic variance without truly disentangling verbosity from content quality. This motivates our approach of counterfactual data augmentation, which disentangles the effect of content and length on reward in a principled manner.

Pearl's Causal Hierarchy The Pearl's Causal Hierarchy (PCH) (Bareinboim et al. 2022; Pearl and Mackenzie 2018) organizes causal reasoning into three hierarchical levels—associational, interventional, and counterfactual—each corresponding to a distinct type of question one can pose about the world. These levels align with fundamental modes of reasoning: observing, acting, and imagining. The first level, *association*, is based on statistical correlations observed in data, typically expressed as conditional probabilities such as $P(y|x)$; for example, one may ask: "How does belief in a disease change when a particular symptom is observed?" The second level, *intervention*, concerns the effects of actions or manipulations, often represented as $P(y|do(x))$ or $P(y_x)$, addressing questions such as: "Will the headache subside if the patient is given the drug?" The third level, *counterfactual*, involves reasoning about alternate outcomes under hypothetical scenarios. Such questions are formulated as $P(y_x|x', y')$, asking, for instance: "If the patient had taken the drug and the headache disappeared, would the headache still have persisted had they not taken the drug?" This hierarchy highlights that higher-level causal reasoning requires stronger assumptions and richer models to support counterfactual inference. See Appendix B for additional examples and formalizations.



(a) Original (b) Ours

Figure 1: Comparison of original perspective and our perspective on reward modeling process.

Structure of Reward Model Common perspectives on reward modeling can be represented by the causal structure in Fig. 1 (a), where a response T is generated from a prompt X and then passed to a reward model which outputs a score R. While the diagram shows a single response for simplicity, RLHF training is conducted on pairwise preference data, applying this structure to both responses in a comparison.[2]

Crucially, however, the reward signal R is not directly supervised from the environment. Although the reward model outputs a scalar value, this value is not grounded in explicit feedback signals (as in traditional RL), but is instead inferred from binary preference labels over response pairs, either annotated by humans or provided by automated judges. As a result, the reward model is trained via comparative supervision: it learns to assign higher scores to preferred responses within triplets of the form $(X, T_{\text{chosen}}, T_{\text{rejected}})$. Over time, this process encourages the model to approximate a reward function that aligns with observed preferences.

3 Causal Interpretation of Length Bias

Length Bias as a Causal Problem Length bias refers to the tendency to assign higher scores to longer responses, even when they are no more informative than shorter ones. This phenomenon frequently arises during RLHF reward model training due to the entanglement between semantic content and response length, making it difficult for the reward model to determine whether its preference stems from content or length and leading it to treat verbosity as a dominant reward signal. From a causal perspective, this can be illustrated by the structure in Fig. 1(b), where each response T is generated from two factors: latent semantic content C and response length L. These two factors may interact, affecting each other, as responses are shaped both by what is said and how extensively it is conveyed.

Motivation for Counterfactuals Since content and length often co-vary in natural data, conventional observational comparisons—such as randomly sampled response pairs from RLHF datasets—are insufficient to isolate the causal effect of length on reward. To address this, we introduce a *counterfactual data augmentation* strategy, which generates synthetic response pairs to answer questions that natural data

[1]Extended version including full technical appendices is available at: https://arxiv.org/abs/2511.12573.

[2]The final preference comparison between reward scores is omitted in the figure to focus on the causal pathways from the prompt X to each response T.

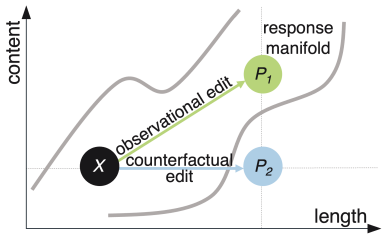


Figure 2: Response dimension over an observation edit (P_1) and a counterfactual (P_2).

cannot support, such as: "What would the reward have been if the response length had been different?"

Fig. 2 illustrates the need for counterfactual data augmentation geometrically. Natural responses lie on a low-dimensional manifold embedded in a space defined by content and length (Appendix C). Because these two attributes are often entangled in real-world data, changing one attribute (e.g., length) typically induces a shift in the other (e.g., content). For instance, an *observational edit* moves along the response manifold—from X to P_1—by varying length, but often alters content implicitly. This is analogous to editing an image: adjusting hairstyle may unintentionally shift perceived gender when the underlying latent representation does not cleanly separate the two. To isolate the effect of a single factor, we must vary one while holding the other fixed. This motivates *counterfactual edits*, which simulate responses *off* the natural manifold (e.g., $X \rightarrow P_2$)—something not achievable with observational data in the response manifold.

Feasibility of Counterfactuals A natural question arises: how can one generate counterfactual data, given that counterfactuals by definition pertain to alternative outcomes that did not actually occur? In our case, such construction is feasible because the setting itself falls under the category of *realizable counterfactuals* (Raghavan and Bareinboim 2025), permitting targeted interventions (Level 2 in PCH) to approximate counterfactual outcomes (Level 3).[3] We can therefore construct counterfactual responses $T_{\tilde{c},\ell'}$ by independently intervening on content and length: setting $C \leftarrow \tilde{c}$ via semantic transplantation and $L \leftarrow \ell'$ via prompt-level control.

When reward models are repeatedly trained on counterfactual preference pairs in which both responses have identical lengths, any residual preference signal must be attributed to semantic differences. This repeated exposure encourages the model to ground its judgments in content alone, gradually attenuating sensitivity to verbosity. In effect, this decouples response length from reward estimation without impairing semantic content of the response. As a result, the influence of length L on reward diminishes, allowing the model to better reflect content-driven distinctions. For a formal treatment of this mechanism, see Appendix E.

[3]Raghavan and Bareinboim (2025) formally show that an L3 distribution is realizable if and only if the target variables (e.g., content C and length L) are not simultaneously subject to conflicting interventions on the same causal parent. See Appendix D for details.

Our augmentation strategy relies on three key assumptions: (1) Each response T can be approximately decomposed into two factors: latent content C and observable length L, such that their influence on the reward is fully mediated through the response; (2) It is feasible to generate alternative responses that modify one factor (e.g., verbosity) while preserving the other (e.g., semantic content), enabling controlled interventions; (3) When two responses have approximately the same length, any difference in model preference is assumed to reflect differences in semantic content quality. These assumptions allow us to treat counterfactual comparisons at fixed length as valid supervision signals for training content-aware reward models.

3.1 Operational Definitions of Length and Content

To enable quantifiable control required for counterfactual data augmentation, we define *length* and *content* in operational terms that allow consistent measurement and manipulation. For *length*, we partition the empirical token distribution of responses into five quantile-based bins—Very Short, Short, Medium, Long, and Very Long—treating responses within the same bin as approximately equal in length. For *content*, which is a latent property, we define it through the relational criterion of semantic equivalence. This relational definition allows us to systematically distinguish between *fixed content* and *varying content* pairs for counterfactual construction. *Fixed Content*: Responses that convey the same meaning despite variations in tone, redundancy, or structure. *Varying Content*: Responses that differ in factuality, specificity, or intent while keeping the length fixed.

4 Length Bias Mitigation Pipeline

To implement our causal approach, we introduce a three-stage framework summarized in Fig. 3. The process consists of three main stages: (1) Counterfactual Data Augmentation: Generate augmented response variants via controlled manipulation on either response length or semantic content, keeping the other approximately constant. (2) Bias Diagnosis: Identify length-driven preference flips by applying targeted length interventions while preserving content. (3) Bias Mitigation: Retrain the reward model using curated counterfactuals that isolate semantic content from stylistic factors like verbosity.

4.1 Counterfactual Data Augmentation Implementations

To disentangle the effect of length on reward, we generate counterfactually augmented response pairs by manipulating either content or length while keeping the other factor approximately fixed. These augmentations produce response pairs aligned in length, enabling controlled comparisons under fixed-length conditions during reward learning. We employ two complementary augmentation strategies[4]:

- *Length-fixed augmentation*: Given a target response, we generate alternative responses that vary in factuality or informativeness by intervening on the semantic content while keeping the length approximately fixed. To achieve

[4]Full augmentation procedures are provided in Appendix F.

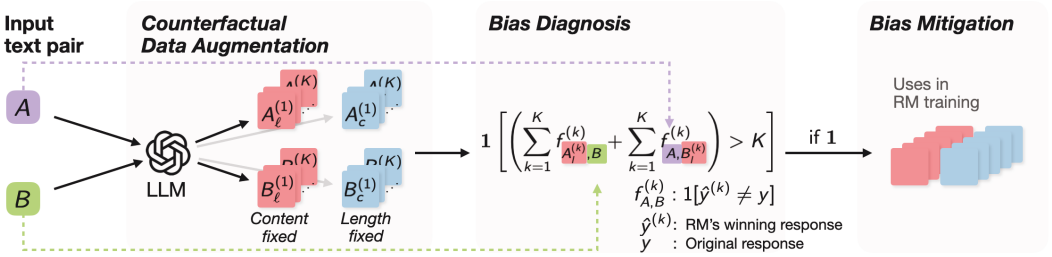


Figure 3: Overview of our method. We generate length- and content-fixed counterfactuals and use them for bias diagnosis and reward model training.

Content Quality	Shorter	Same	Longer
Better	✓	✓	✓
Same	✗	✗	length bias
Worse	✗	length bias	length bias

Table 1: Rules for diagnosing length bias based on the winning response's content quality and relative length: the model's preference is acceptable (✓), implausible (✗), or likely indicative of length bias (length bias).

this, we apply transformations such as detail removal, elaboration, information substitution, or rewriting figurative expressions into literal descriptions.

- *Content-fixed augmentation*: Given a target response, we generate alternatives that vary in verbosity by intervening on length while preserving the original meaning. Here, the length is aligned to that of the reference response used to determine the original preference label. To preserve semantic content while varying length, we apply surface-level augmentations such as filler insertion (deletion), pleonasm (simplification), redundant sentence reusing (pruning), paraphrasing, and format changes.

After generating responses, it is essential to verify that the intended factor has been successfully manipulated. As described in Sec. 3.1, length is operationalized via token-level binning, enabling automatic verification through bin membership. In contrast, content is a latent property that cannot be reliably assessed using simple heuristics. To ensure fidelity, we apply automated editing strategies followed by semantic filtering using a binary classifier that checks whether semantic content is preserved. These verification steps confirm whether the intended manipulation was successful and enhance the overall quality of the augmented data.

4.2 Diagnosing Length Bias

We define a diagnostic rule table (Table 1) that categorizes each case based on the relative content quality and length of the winning response, and determines whether the model's preference is attributable to content or likely influenced by length bias. According to this scheme, length bias is diagnosed when the model prefers a longer response with worse content, excluding ties from both diagnosis and training.

Preference flips. To scale up the diagnostic rules in Table 1 for large-scale evaluation, we introduce a binary decision rule based on *preference flips*, which are induced by controlled length interventions using content-fixed augmented responses. A preference flip occurs when the reward model reverses its ranking of a response pair solely due to a change in length, with semantic content held constant. Let $R(X, T)$ denote the reward score assigned to response T given prompt X. Suppose the original model preference is $R(X, A) > R(X, B)$. A preference flip is said to occur if, for a counterfactual variant A' or B' with the same content but altered length, the ranking reverses—e.g., $R(X, A') < R(X, B)$ or $R(X, A) < R(X, B')$. When the model initially prefers the longer response but reverses its choice after a content-preserving length adjustment, we interpret this as evidence that the original preference was driven by verbosity rather than semantic quality.

For each original preference pair (A, B), we generate K content-fixed length variants and evaluate the model's consistency against them (e.g., A vs B'_1, A vs B'_2, etc.). A pair is considered biased if the number of preference flips exceeds half of the number of counterfactuals. Formally, we apply the indicator function $\mathbf{1}[(\sum_{k=1}^{K} f_{A_l}^{(k)} + \sum_{k=1}^{K} f_{B_l}^{(k)}) > K]$, where $f_{A_l}^{(k)} = \mathbf{1}[\hat{y}^{(k)} \neq y]$ denotes a flipped preference under the k-th intervention on response A, and y is the original model preference label—i.e., the response that was favored in the original (A, B) pair. This criterion detects cases where length changes alone consistently alter model decisions.

To generalize this diagnosis, we define the *flip ratio* F:

$$F_{(A,B)} = \frac{\#\text{ of flipped preferences}}{\text{Total counterfactual comparisons}}.$$

Pairs with $F > 0.5$ are flagged as length-biased. This continuous metric enables scalable and automated diagnostics aligned with the rule-based intuition, while providing fine-grained control for downstream filtering and training.

Final counterfactual data used for bias mitigation. After identifying bias-prone examples via preference flips, we selectively use counterfactually augmented response pairs for reward model training. For response pairs that exhibit length bias—i.e., more than half of the content-fixed counterfactual pairs result in preference flips—we include the flipped examples in the training set. In addition, we incorporate the

Figure 4: Experimental pipeline with edge styles indicating flow type. Purple nodes represent data transformation stages; orange nodes algorithmic processing steps. Black arrows data transformation; red dashed arrows model-based processing or state change.

length-fixed variants of the original responses to further disentangle length from content during reward learning.

4.3 Mitigating Length Bias

To illustrate how the selected counterfactually augmented data mitigates length bias, consider a response pair (A, B) that has been diagnosed as length-biased through content-fixed augmentation. In the original preference data, the model favored A due to its verbosity rather than its semantic quality, meaning the original supervision signal does not reflect a true preference. To correct this, we form a new training pair by combining the content-fixed counterfactual A'—which preserves the meaning of A while matching the length of B—with the original response B, forming the pair (A', B). In this counterfactual pair, where length is neutralized, we revise the supervision to favor B, yielding a more accurate learning signal grounded in semantic content.

In addition to the flipped preference pairs from content-fixed augmentations, we also incorporate length-fixed augmentations to directly supervise semantic quality. For each response A, we use its pre-generated length-fixed variant A'', which has degraded semantic content but maintains length, and train the model to prefer A over A''. This setup encourages the reward model to distinguish fine-grained semantic differences under fixed stylistic conditions, enabling learning signals that reflect content alone. For instance, if A'' is semantically inferior to A, and A is determined to be worse than B based on the content-controlled pair (A', B), the model learns a content-grounded ranking: $A'' < A' = A < B$.

These counterfactual augmentations ensure that the reward model learns to prioritize semantic content over superficial stylistic features such as verbosity by providing targeted supervision—via content- and length-fixed variants—that explicitly disentangles content quality from length artifacts.

5 Experiments

We present an overview of our experimental pipeline in Fig. 4, which summarizes the three core stages of our method: counterfactual data augmentation, bias diagnosis, and bias mitigation through reward model fine-tuning.

5.1 Data Augmentation

We use the RLHF preference dataset from RLHFlow (Dong et al. 2024), consisting of 699k prompt-response pairs with pairwise preference labels from seven sources, for its large scale and diverse annotation sources. For augmentation, we selected **GPT-4o-mini** (OpenAI 2024) due to its strong semantic fidelity and stylistic control.

To study length bias, we filtered for examples where the preferred response is longer and the two responses fall into different length bins (Appendix G), discarding pairs with extreme disparities ($\geq$4 bins apart). This yielded 225,358 examples, from which we randomly sampled 50,000 for augmentation. Using controlled editing strategies,[5] we generated 474k content-fixed and 471k length-fixed response pairs, totaling approximately 945k augmented comparisons—a 19$\times$ increase over the original sample.

Stage	Content	Length
Augmented pairs (pre-filtering)	474k	471k
Filtered pairs	473k	466k
Length bias pairs	199k	214k

Table 2: Summary of augmentation and filtering statistics.

To verify whether the intended factor (content or length) is correctly preserved, we fine-tuned a binary classifier based on `all-mpnet-base-v2` (Song et al. 2020).[6] After filtering, 472k content-fixed and 466k length-fixed response pairs were retained. These filtered counterfactuals form a reliable basis for diagnosing length bias.

5.2 Length Bias Identification and Mitigation Data Construction

Length bias identification. To measure the presence of length bias, we remove all original preference labels from the RLHFlow dataset and re-score each prompt–response pair using a reference reward model, `OpenLLaMA-3B` (Geng and Liu 2023). Then, we conducted content-fixed comparison using counterfactual responses that preserve semantic content while varying length. A *flip* is recorded when a model's preference reverses due to a change in length alone. Among 49,861 pairs, 23,651 (47.43%) exhibited length bias. See Appendix J for full distribution.

Mitigation data construction. For each pair classified as length-biased, we construct a mitigation dataset by combining content-fixed and length-fixed augmentations. From the content-fixed set, we retain only those that cause a preference flip, ensuring the bias is empirically observed. Then, for each confirmed flip, we include the corresponding length-fixed augmentations to reinforce content sensitivity. This yields 198,778 flipped content-fixed pairs and 213,699 aligned length-fixed augmentations. After deduplication, we obtain 412,286 unique (prompt, chosen, rejected) triplets which will

[5]Examples and templates illustrating these augmentation strategies are provided in Appendix H.

[6]We describe implementation details for all models—including, cross-encoders, reward models, and policy models—in Appendix I.

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 3: Accuracy of reward models across three datasets: **RewardBench-1**, **RewardBench-2**, and length-controlled (LC) accuracy from **Chatbot Arena**. (* indicates models trained with our method.)

be used in reward model fine-tuning. The total number of data points processed is summarized in Table 2.

Reward model finetuning. To mitigate length bias, we fine-tune reward models on a counterfactually augmented dataset that disentangles verbosity from semantic content. We utilized two baseline models to fine-tune: `OpenLLaMA-3B` and its RLHF variant reward model, `HH-RLHF_RM_OpenLLaMA-3B` (Diao et al. 2024; Dong et al. 2023).

Fine-tuned reward models. We evaluate five reward models: (1) **HRO**, the baseline reward model `HH-RLHF_RM_OpenLLaMA-3B`; (2) **ODIN** (Chen et al. 2024), a recent method which mitigates length bias via dual-head reward modeling. Although the original ODIN used Vicuna-7B, we reimplemented it on the `OpenLLaMA-3B` backbone to eliminate confounding effects from base model performance differences. This ensures that any observed differences in evaluation are attributable to methodological differences rather than disparities in model capacity or pretraining quality. (3) **CDA_OpenLM**, a reward model obtained through fine-tuning `OpenLLaMA-3B` on our counterfactually augmented dataset; (4) **CDA_LoRA**, a LoRA-based fine-tuning of `HRO` using our mitigation data; and (5) **CDA_HRO**, a full fine-tuning of `HRO` on the same dataset. Comprehensive details of all evaluation experiments, including dataset specifications and repeated-run statistics, are provided in Appendix K.

Evaluation of reward model length bias reduction. We evaluate reward models using two complementary metrics:

- **RewardBench Average Score**: RewardBench (Lambert et al. 2025; Malik et al. 2025) provides a comprehensive evaluation of general reward model performance aligned with human preferences. We use both versions of the benchmark, with version 2 being more challenging than version 1. For consistency with our length-bias diagnosis and mitigation setup, we exclude tie cases from RewardBench-2 in our reporting.
- **Length-controlled accuracy**: This metric directly measures the extent to which reward models rely on verbosity by evaluating whether they correctly prefer shorter responses when appropriate. Using Chatbot Arena pairwise preferences (Chiang et al. 2024), we select pairs in which the preferred response is shorter by at least two token-length bins and check whether the reward model assigns higher scores to the concise option.

Figure 5: Reward distribution across response lengths on RewardBench-1 (top) and RewardBench-2 (bottom).

Together, these metrics assess whether reducing length bias compromises general reward model performance.

Table 3 reports category-wise and mean accuracy on RewardBench along with length-controlled accuracy. On RewardBench-1, `CDA_HRO` achieves the highest mean accuracy, and both `CDA_HRO` and `CDA_OpenLM` exhibit more balanced category-wise performance, with notable improvements on bias-sensitive subsets (`chat hard`, `safety`, `reasoning`). In contrast, baseline models display more uneven patterns with similar overall scores. For RewardBench-2, `CDA_LoRA` and `HRO` show strong peaks in individual categories, while `CDA_OpenLM` and `CDA_HRO` provide more stable performance. As RewardBench-2 is more challenging and less sensitive to verbosity, large gains are not expected, and our models perform comparably to the baseline. This trend is consistent with the length-controlled evaluation, where `CDA_OpenLM` and `CDA_HRO` achieve substantially higher LC accuracy than the baseline, demonstrating robustness when length cannot be used as a cue. Although `CDA_LoRA` remains competitive overall, its lower LC accuracy (24.80%) reflects the limitations of partial fine-tuning in mitigating length bias. Across both RewardBench versions, our CDA-based models maintain competitive overall accuracy with more stable category-wise behavior, while substantially improving length-controlled accuracy.

Taken together, these results show that our approach mitigates length bias without compromising general reward model performance, overcoming the trade-off commonly observed in baseline methods.

Reward distribution across length. To examine how reward models handle verbosity, we visualize reward–length distributions for RewardBench-1 and the more challenging RewardBench-2 (Fig. 5). In both datasets, `HRO` shows a

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

Table 4: Length controlled winrate of RLHF models on AlpacaEval. (* for ours.)

strong positive correlation between response length and reward, revealing substantial length bias. In contrast, our counterfactually fine-tuned models (CDA_HRO, CDA_OpenLM) and ODIN yield more vertically aligned distributions, indicating reduced sensitivity to verbosity. The gap becomes even clearer on the harder RewardBench-2, where the baseline's bias intensifies while our models remain robust.

Policy model finetuning. To evaluate the downstream impact of length bias mitigation, we fine-tune a policy via Supervised Fine-Tuning (SFT) followed by PPO (Schulman et al. 2017), using reward models with and without mitigation. This allows us to assess how improvements in reward modeling affect final policy behavior.

We evaluate six RLHF policy models, all initialized from the same SFT model based on OpenLLaMA-3B[7]: (1) **OpenLM**, the unaligned base OpenLLaMA-3B model; (2) **SFT**, trained on supervised instruction-following data; (3) **PPO_HRO**, fine-tuned with the baseline reward model HRO; (4) **ODIN**, trained using the ODIN reward model reimplemented on OpenLLaMA-3B; (5) **PPO_CDA_OpenLM**, trained with our counterfactual reward model CDA_OpenLM; and (6) **PPO_CDA_HRO**, trained with CDA_HRO. This setup enables controlled comparison of reward strategies and the effectiveness of counterfactual data.

Final RLHF performance. We follow the AlpacaEval protocol (Dubois, Liang, and Hashimoto 2024), where models are evaluated against LLaMA-2-7B-chat-hf (Touvron et al. 2023). Since our base policy model is OpenLLaMA-3B, which is derived from LLaMA-1, we select LLaMA-2-7B-chat as a reference—upgrading both the version (LLaMA-1 to LLaMA-2) and model size (3B to 7B). This choice ensures a stronger and more up-to-date judge model for comparison. Additional evaluations are in Appendix L.

Table 4 shows that our RLHF-trained model, PPO_CDA_HRO, achieves the highest length-controlled winrate (37.18%), more than doubling that of PPO_HRO (18.97%) and SFT (16.97%). While baseline models achieve reasonable overall winrates, their sharp drop under length control indicates reliance on verbosity. In contrast, our CDA-trained policies produce shorter yet higher-quality responses, demonstrating that gains at the reward-model level carry over to downstream RLHF behavior. Unlike ODIN, which reduces length at the cost of winrate, our method improves conciseness without degrading performance.

[7]Note that we exclude CDA_LoRA from PPO training, as parameter-efficient fine-tuning under frozen reward heads showed limited effectiveness in mitigating bias (Table 3).

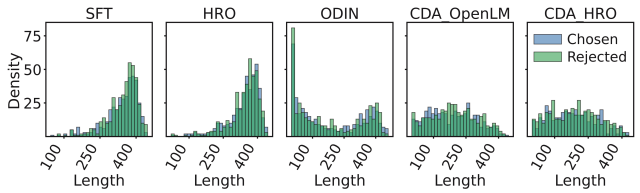


Figure 6: Length distribution of model outputs. OpenLM is excluded from this comparison as it is an untrained baseline model without any supervised fine-tuning.

To identify factors driving win rate differences in AlpacaEval, we analyzed response-length distributions. As shown in Figure 6, traditional reward models like SFT and PPO_HRO, tend to favor longer responses, with a noticeable skew toward higher word counts. In contrast, models trained with our counterfactual data augmentation approach (PPO_CDA_OpenLM and PPO_CDA_HRO) show a more balanced distribution, with shorter, content-rich responses being selected more frequently. This indicates that our improved performance stems from the ability to generate more concise, high-quality outputs rather than simply producing longer responses.

Summary. Taken together, our results show that counterfactual reward-model fine-tuning improves robustness to length bias without sacrificing overall performance. Across reward-level benchmarks and downstream RLHF evaluation, CDA-based models avoid verbosity-driven preferences and enable policies to generate more concise, content-faithful responses, outperforming prior approaches.

6 Conclusion

We presented a causal framework for mitigating length bias in reward models trained through reinforcement learning from human feedback (RLHF). Our method uses counterfactual data augmentation to disentangle the effect of semantic content from verbosity, enabling reward models to better reflect human preferences grounded in meaning rather than surface features. By generating content-fixed and length-fixed response pairs, we allow the reward model to learn preferences that are robust to stylistic confounds and focused on underlying semantic quality. Empirical evaluations demonstrate that our approach substantially reduces length-driven preference errors while maintaining or improving alignment performance across standard RLHF benchmarks. In particular, models trained with our counterfactually augmented data consistently prefer more concise yet informative responses and show greater robustness to stylistic variations. Moreover, downstream policy models optimized with these improved reward signals generate responses that are not only shorter on average but also rated higher in informativeness and relevance. These findings underscore the utility of causality in enhancing the fidelity and interpretability of reward models. While our method assumes a clean separation between content and length, it opens the door for future extensions that address additional confounding factors—such as tone, coherence, or factuality—through appropriate expansions of the underlying causal graph.

Acknowledgments

We thank anonymous reviewers for constructive comments to improve the manuscript. This work was supported by NRF (RS-2023-00211904/50%, RS-2023-00222663/50%) grant funded by the Korean government.

References

Bareinboim, E.; Correa, J.; Ibeling, D.; and Icard, T. 2022. On Pearl's Hierarchy and the Foundations of Causal Inference (1st edition). In Geffner, H.; Dechter, R.; and Halpern, J., eds., *Probabilistic and Causal Inference: the Works of Judea Pearl*, 507–556. ACM Books.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cai, J.; Zhu, J.; Sun, R.; Wang, Y.; Li, L.; Zhou, W.; and Li, H. 2025. Disentangling length bias in preference learning via response-conditioned modeling. *arXiv preprint arXiv:2502.00814*.

Chen, L.; Zhu, C.; Chen, J.; Soselia, D.; Zhou, T.; Goldstein, T.; Huang, H.; Shoeybi, M.; and Catanzaro, B. 2024. ODIN: Disentangled Reward Mitigates Hacking in RLHF. In *Forty-first International Conference on Machine Learning*.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhu, B.; Zhang, H.; Jordan, M.; Gonzalez, J. E.; et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113.

Diao, S.; Pan, R.; Dong, H.; Shum, K.; Zhang, J.; Xiong, W.; and Zhang, T. 2024. Lmflow: An extensible toolkit for finetuning and inference of large foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, 116–127.

Dong, H.; Xiong, W.; Goyal, D.; Zhang, Y.; Chow, W.; Pan, R.; Diao, S.; Zhang, J.; SHUM, K.; and Zhang, T. 2023. RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment. *Transactions on Machine Learning Research*.

Dong, H.; Xiong, W.; Pang, B.; Wang, H.; Zhao, H.; Zhou, Y.; Jiang, N.; Sahoo, D.; Xiong, C.; and Zhang, T. 2024. RLHF Workflow: From Reward Modeling to Online RLHF. *Transactions on Machine Learning Research*.

Dubois, Y.; Liang, P.; and Hashimoto, T. 2024. Length-Controlled AlpacaEval: A Simple Debiasing of Automatic Evaluators. In *First Conference on Language Modeling*.

Geng, X.; and Liu, H. 2023. Open LLaMA: An Open Reproduction of LLaMA. https://github.com/openlm-research/open_llama.

Huang, Z.; Qiu, Z.; Wang, Z.; Ponti, E.; and Titov, I. 2025. Post-hoc Reward Calibration: A Case Study on Length Bias. In *The Thirteenth International Conference on Learning Representations*.

Kaushik, D.; Hovy, E.; and Lipton, Z. 2020. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *International Conference on Learning Representations*.

Kim, H.; Oh, S.; and Lee, S. 2025. Mitigating Length Bias in RLHF through a Causal Lens. arXiv:2511.12573.

Lambert, N.; Pyatkin, V.; Morrison, J.; Miranda, L. J. V.; Lin, B. Y.; Chandu, K.; Dziri, N.; Kumar, S.; Zick, T.; Choi, Y.; et al. 2025. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 1755–1797.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C.; Manning, C. D.; Ré, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; Wang, J.; Santhanam, K.; Orr, L.; Zheng, L.; Yuksekgonul, M.; Suzgun, M.; Kim, N.; Guha, N.; Chatterji, N.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2023. Holistic Evaluation of Language Models. arXiv:2211.09110.

Liu, T.; Xiong, W.; Ren, J.; Chen, L.; Wu, J.; Joshi, R.; Gao, Y.; Shen, J.; Qin, Z.; Yu, T.; Sohn, D.; Makarova, A.; Liu, J. Z.; Liu, Y.; Piot, B.; Ittycheriah, A.; Kumar, A.; and Saleh, M. 2025. RRM: Robust Reward Model Training Mitigates Reward Hacking. In *The Thirteenth International Conference on Learning Representations*.

Malik, S.; Pyatkin, V.; Land, S.; Morrison, J.; Smith, N. A.; Hajishirzi, H.; and Lambert, N. 2025. RewardBench 2: Advancing Reward Model Evaluation. *arXiv preprint arXiv:2506.01937*.

OpenAI. 2024. GPT-4o: OpenAI's new multimodal model. https://openai.com/index/gpt-4o. Accessed: 2025-04-01.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Pearl, J.; and Mackenzie, D. 2018. *The Book of Why: The New Science of Cause and Effect*. New York, NY: Basic Books. ISBN 9780465097609. First Edition.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Raghavan, A.; and Bareinboim, E. 2025. Counterfactual Realizability. In *The Thirteenth International Conference on Learning Representations*.

Saito, K.; Wachi, A.; Wataoka, K.; and Akimoto, Y. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shen, W.; Zheng, R.; Zhan, W.; Zhao, J.; Dou, S.; Gui, T.; Zhang, Q.; and Huang, X.-J. 2023. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2859–2873.

Singhal, P.; Goyal, T.; Xu, J.; and Durrett, G. 2024. A Long Way to Go: Investigating Length Correlations in RLHF. In *First Conference on Language Modeling*.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33: 3008–3021.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, C.; Zhao, Z.; Jiang, Y.; Chen, Z.; Zhu, C.; Chen, Y.; Liu, J.; Zhang, L.; Fan, X.; Ma, H.; and Wang, S. 2025. Beyond Reward Hacking: Causal Rewards for Large Language Model Alignment. arXiv:2501.09620.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.