Beyond Excess and Deficiency: Adaptive Length Bias Mitigation in Reward Models for RLHF

Yuyan Bu[*], **Liangyu Huo**[*][†], **Yi Jing**, **Qing Yang**
Du Xiaoman (Beijing) Science Technology Co., Ltd.
{buyuyan, huoliangyu, jingji, yangqing}@duxiaoman.com

Abstract

Reinforcement learning from human Feedback (RLHF) is crucial for aligning large language models (LLMs) with human values. However, it has been noted that reward models in RLHF often exhibit unintended biases, such as an overemphasis on response length based on the erroneous assumption that longer responses are universally preferred. This “length bias” can lead to excessively verbose responses that compromise the quality of LLMs alignment. Previous efforts to mitigate length bias in reward models have inadvertently decreased their accuracy by neglecting the legitimate influence of response length on human preferences. In this work, we argue that response length is a context-specific factor in human evaluations, with different queries naturally eliciting varying preferences for response length. We propose an adaptive approach to modeling length preference that dynamically adjusts the influence of response length in reward evaluations according to the context of the query. Experimental results demonstrate that our adaptive approach effectively balances the mitigation of undesired length hacking and alignment accuracy, reducing unnecessary verbosity while improving overall response quality.

1 Introduction

> *“Excess and deficiency are equally at fault.”*
> — Confucius

Reinforcement Learning from Human Feedback (RLHF)(Ouyang et al., 2022; Bai et al., 2022) has played a pivotal role in the impressive advancements of state-of-the-art large language models (LLMs)(Achiam et al., 2023; Team et al., 2023; Anthropic, 2023). The standard RLHF pipeline involves two main stages: first, a reward model (RM) is trained on data that captures human preferences for responses to specific prompts, and second, the language model is optimized to generate responses that maximize the learned reward through reinforcement learning (Ziegler et al., 2019). A key determinant of RLHF’s success is that the RM must genuinely reflect human preferences (Zhuang and Hadfield-Menell, 2020). However, achieving an automatic proxy that perfectly replicates human judgment is challenging in practice (Gao et al., 2023). Reward models often face challenges in generalizing to out-of-distribution data (Eisenstein et al., 2023), which can cause the policy model to maximize the reward score without fully aligning with human intent—an issue known as “reward hacking” (Skalse et al., 2022).

Verbosity, one of the most common reward hacking problems, occurs when LLMs generate excessively long responses to exploit human raters’ preference for detailed content. This length hacking issue has been identified in both explicit (Singhal et al.) and implicit (Park et al., 2024) reward modeling methods. To address this, approaches have been developed to reduce the correlation between response length and reward scores by adjusting data processing (Liu et al., 2024) or model design (Dubois et al., 2024; Shen et al., 2023; Chen et al.). However, while these approaches successfully weaken the association between length and reward scores, they unintentionally reduce the overall accuracy of RMs.

We attribute this drop in accuracy to treating length uniformly across all queries. We argue that response length should be treated as a context-dependent factor in human evaluations, as different queries naturally demand varying lengths. For instance, in open-ended questions like "Explain to me like I’m five," users typically prefer longer, more detailed responses, where length strongly correlates with user intent (referred to as length-sensitive queries). Conversely, conciseness is favored in more specific queries like "What gives non-pepper things like garlic their spice?" as users prefer di-

[*] Equal contribution.
[†] Corresponding author.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 3091–3098
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

rect and to-the-point answers (referred to as length-neutral queries). Existing debiasing approaches tend to suppress the influence of length across both query types uniformly, failing to adapt to the natural variations in length preference. This indiscriminate suppression leads to a decline in alignment performance, particularly for length-sensitive queries where length plays a legitimate role in user satisfaction. We present a detailed illustration in A.1.

To address this issue, we propose an approach that implements Adaptive Length Bias Mitigation (ALBM) in reward models. Specifically, ALBM first decouples length bias from the original reward, and then reintegrates the length reward with the quality reward based on the nature of the query. Experiments show that ALBM outperforms existing debiasing techniques by improving alignment accuracy without inducing verbosity-related reward hacking. Additionally, in the reinforcement learning phase, we observe that the policy model supervised by our debiased RM can generate higher-quality responses compared with vanilla RM.

2 Preliminary

We focus on the widely adopted RLHF pipeline, which consists of three main stages: (1) supervised fine-tuning (SFT); (2) reward modeling; and (3) reinforcement learning optimization. Our primary attention is on the latter two phases.

Reward modeling. Following Touvron et al. (2023), the RM is initialized from an SFT model, appending a randomly initialized linear layer at the end to project the feature representation into a scalar reward value. The RM is trained to minimize the Bradley–Terry loss (Bradley and Terry, 1952) on pair-wise comparisons of model responses:

$$\begin{aligned}\mathcal{L}_{\text{RM}} &= -\mathbb{E}\left[\log \sigma\left(r_\theta(x, y_w) - r_\theta(x, y_l)\right)\right] \\ &= -\mathbb{E}\left[\log \sigma\left(g_\psi \circ f_\phi(x, y_w) - g_\psi \circ f_\phi(x, y_l)\right)\right]\end{aligned}$$

Here, $r_\theta(x, y)$ represents the scalar reward for a given prompt x and response y. y_w and y_l denote the chosen and rejected responses, respectively. $\sigma(\cdot)$ refers to the sigmoid function. The trainable parameters θ consist of the foundational language model f_ϕ and the linear head g_ψ, such that the reward is computed as: $r_\theta(x, y) = g_\psi \circ f_\phi(x, y)$.

Reinforcement learning optimization. We utilize the proximal policy optimization (PPO) algorithm for reinforcement learning optimization. In this process, the RM serves as a proxy for human feedback on the responses generated by the policy during training. The policy parameters w are fine-tuned by maximizing the following objective:

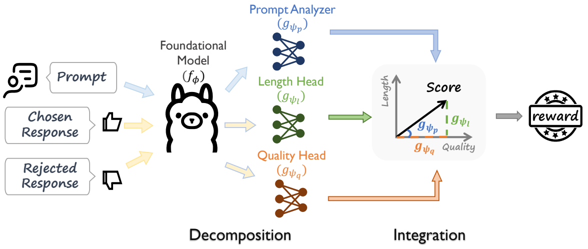

Figure 1: Overview of ALBM. The overall reward is decomposed into length and quality components, then reintegrated according to prompt analysis to obtain the final reward score.

$$\max_{w} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\pi_w}}\left[r_\theta(x, y)\right] - \beta\, \mathbb{D}_{\text{KL}}\left(\pi_w(y \mid x) \,\|\, \pi^{\text{SFT}}(y \mid x)\right)$$

Here, the SFT policy π^{SFT} is used to initialize policy π_w. $\mathcal{D}_{\pi_w}$ represents the prompt-response pairs sampled from π_w. The parameter β controls the strength of the KL-penalty term to prevent the policy π_w from drifting too far from the SFT model.

3 Method

ALBM consists of two key components to account for length bias in preference modeling adaptively. First, we decouple the original reward into two distinct parts: a length reward and a quality reward, the latter capturing human preferences independent of length bias, as discussed in prior work (Chen et al.). By isolating the influence of response length from the overall reward, we prevent the model from exploiting length to artificially boost reward scores, thereby avoiding length-based reward hacking. Next, based on an analysis of the input prompt, we adaptively recombine the length reward with the quality reward to derive the final reward score. This adaptive integration allows the model to consider context-dependent preferences for response length, resulting in improved performance. Figure 1 provides an overview of ALBM.

3.1 Reward Decomposition

Inspired by the Product of Experts (PoE) framework (Singhal et al.), we adopt a disentanglement strategy to isolate the influence of response length. To reduce computational overhead and maximize the utility of the pre-trained language model, we introduce two distinct final linear layers as different "experts" to separate the learning of true human intent from length bias, similar to Chen et al.. To be specific, we append two separate linear heads to the shared backbone f_ϕ: a quality reward head (g_{ψ_q}), which focuses on learning quality-related preferences independent of length, and a length reward head (g_{ψ_l}), which captures preferences influenced by length bias. Each reward head computes its respective reward score for a given prompt-response pair and contributes to the overall reward loss:

$$\begin{aligned}\mathcal{L}_{\texttt{DR}} = &-\mathbb{E}\left[\log(\sigma(g_{\psi_q} \circ f_\phi(x, y_w) - g_{\psi_q} \circ f_\phi(x, y_l)))\right] \\ &-\mathbb{E}\left[\log(\sigma(g_{\psi_l} \circ f_\phi(x, y_w) - g_{\psi_l} \circ f_\phi(x, y_l)))\right]\end{aligned}$$

To ensure that each head specializes in its designated role, we introduce both explicit and implicit constraints. For the explicit constraint, we design a loss function that enhances the correlation between the length reward and the response length while minimizing the correlation between the quality reward and the response length. Specifically, we use the following loss function:

$$\begin{aligned}\mathcal{L}_{\texttt{EL}} = &\left|\rho\left(g_{\psi_q} \circ f_\phi(x, y), L(y)\right)\right| \\ &-\rho\left(g_{\psi_l} \circ f_\phi(x, y), L(y)\right)\end{aligned}$$

Here, $L(y)$ represents the number of tokens in the response y, and $\rho(X, Y)$ denotes the Pearson correlation between X and Y, computed over the global minibatch. To further strengthen the disentanglement, we impose an implicit constraint by enforcing orthogonality between the projection weights of the two heads as follows:

$$\mathcal{L}_{\texttt{IL}} = \left|\mathbf{W}_{\psi_q} \mathbf{W}_{\psi_l}^T\right|$$

$\mathbf{W}_{\psi_q}$, $\mathbf{W}_{\psi_l}$ represent the linear projection weights for the quality and length rewards, respectively.

3.2 Adatively Utilization of Length Preferance

We dynamically adjust the influence of length bias in our reward modeling by introducing an additional head, g_{ψ_p}, which analyzes the input prompt x. This prompt analyzer learns to predict how much the response length should influence the overall reward based on the prompt's content. By capturing this relationship, g_{ψ_p} guides the appropriate weighting of the length reward when computing the total reward: $g_{\psi_q} \circ f_\phi(x, y) + (g_{\psi_p} \circ f_\phi(x)) \circ g_{\psi_l} \circ f_\phi(x, y)$. As a result, the primary ranking loss used to train the RM becomes as follows:

$$\begin{aligned}\mathcal{L}_{\texttt{R}} = -\mathbb{E}\Big[\log\Big(\sigma\Big(&g_{\psi_q} \circ f_\phi(x, y_w) \\ &+(g_{\psi_p} \circ f_\phi(x)) \circ g_{\psi_l} \circ f_\phi(x, y_w)\Big) \\ &-g_{\psi_q} \circ f_\phi(x, y_l) \\ &-(g_{\psi_p} \circ f_\phi(x)) \circ g_{\psi_l} \circ f_\phi(x, y_l)\Big)\Big]\end{aligned}$$

Table 1: Performance comparison of RMs trained with different methods. ACC: Accuracy. CORR: Spearman correlation between reward scores and response length. LN-FR: Failure rate on length-neutral data. LS-FR: Failure rate on length-sensitive data.

Methods	ACC(↑)	CORR	LN-FR(↓)	LS-FR(↓)
Vanilla	0.6223	0.5105	0.6416	0.1945
Bal	0.5906	-0.1067	0.3721	0.4431
Odin	0.5792	-0.0670	0.3905	0.4508
ALBM	**0.6318**	-0.0209	0.4655	0.3049

We conduct training with g_{ψ_q},g_{ψ_l} and g_{ψ_p} to minimize the final loss:

$$\mathcal{L}_{final} = \mathcal{L}_{\texttt{R}} + \lambda_{\texttt{DR}}\mathcal{L}_{\texttt{DR}} + \lambda_{\texttt{EL}}\mathcal{L}_{\texttt{EL}} + \lambda_{\texttt{IL}}\mathcal{L}_{\texttt{IL}}$$

where $\lambda_{\texttt{DR}}, \lambda_{\texttt{EL}}, \lambda_{\texttt{IL}}$ are regularization coefficients.

4 Experiments

4.1 Experimental Setup

Our experiments are primarily conducted on the WebGPT dataset (Nakano et al., 2021) using Vicuna-7B (Zheng et al., 2023) as the foundation model. For a fine-grained analysis, the data is further divided into a length-sensitive subset and a length-neutral subset, based on the length relationship between the chosen and rejected responses. Moreover, for out-of-distribution generalization analysis, we evaluate our approach on the Stack (Lambert et al., 2023) and RM-static (Bai et al., 2022) datasets. To assess performance across different base models, we conduct experiments on Vicuna-13B and LLAMA3-8B (AI@Meta, 2024). We compare our approach with two typical baselines for mitigating length hacking: a data intervention method (Bal, Singhal et al.) and a model intervention method (Odin, Chen et al.). More detailed descriptions can be found in A.3.

4.2 Performance of Reward Modeling

Table 1 compares the performance of RMs trained with different methods. The vanilla RM exhibits a strong correlation between reward scores and response length, performing poorly on length-neutral data. Existing debiasing methods, such as Bal and Odin, effectively reduce this correlation, improving performance on length-neutral data. However, they tend to overcorrect, excessively suppressing length bias, leading to poor performance on length-sensitive data and a decline in alignment accuracy. In contrast, our ALBM approach strikes a better balance by reducing the length-score correlation while preserving an appropriate level of length bias.

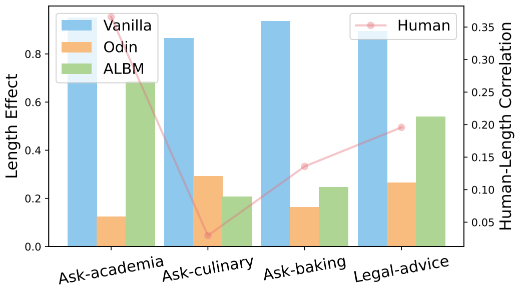

Figure 2: Adaptive Length Bias Utilization Analysis. Compared to existing approaches, Length Effect of ALBM aligns more closely with trends in humans.

This leads to performance improvements on length-neutral data, with only minimal degradation on length-sensitive data, resulting in a substantial overall enhancement in alignment accuracy compared to the baselines. The generalization analysis across different datasets and models is provided in A.4. In addition to better performance, the approach adds minimal computational overhead, with the forward pass time increasing by just 0.75% compared to a standard reward model.

4.3 Adaptive Utilization of Length Bias

To validate the adaptive utilization of length bias in our approach, we conducted an analysis using randomly sampled data from various categories within the SHP dataset (Ethayarajh et al., 2022). We introduced the Length Effect metric to quantify the model's reliance on response length during preference judgments:

$$Length\ Effect = |Acc(l\text{-}c, s\text{-}r) - Acc(s\text{-}c, \text{l-r})|$$

This metric captures the asymmetry of length bias—if the RM favors longer responses, increasing the length of the chosen response (lengthened chosen, l-c) while shortening the rejected response (shortened rejected, s-r) will lead to higher accuracy, and the reverse will reduce accuracy. The more significant this discrepancy, the stronger the Length Effect, indicating that the model's judgments are more influenced by response length. For each category, we also computed the correlation between human ratings and response length as a reference. As shown in Figure 2, the vanilla RM consistently exhibits a higher Length Effect across categories, whereas the debiasing baseline Odin shows a lower Length Effect. In contrast, ALBM adaptively utilizes length bias across different categories, aligning more closely with human preferences.

Table 2: Average length of generated response.

Model	SFT	Vanilla	Bal	Odin	ALBM
Length	198	261	125	206	228

Table 3: Win rates of various models against the vanilla model. Evaluated by GPT-4o.

	All			Length-Sensitive Data			Length-Neutral Data		
Model	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
SFT	0.16	0.34	0.50	0.12	0.36	0.52	0.20	0.32	0.48
Bal	0.25	0.30	0.45	0.20	0.24	0.56	0.30	0.36	0.34
Odin	0.28	0.41	0.31	0.26	0.34	0.40	0.30	0.48	0.22
ALBM	0.37	0.35	0.28	0.40	0.30	0.30	0.34	0.4	0.26

4.4 Impact on Downstream RL Optimization

We further evaluated the proposed approach by assessing the performance of the aligned policies after RL training. To ensure a fair comparison, we selected checkpoints at the same training step in the convergence stage for each approach. As shown in Table 2, the policy trained with vanilla RM generates longer responses than the SFT model, indicating its susceptibility to length bias. All debiasing methods effectively reduce the length of generated responses. Notably, the data-intervention method (Bal) achieves the most significant reduction, even generating responses shorter than the SFT model. While the model-intervention methods (Odin and ALBM) also reduce response length relative to the vanilla model, they still yield slightly longer responses than the SFT model.

To evaluate the quality of the generated responses, we compared the win rates of models trained with different debiased RMs against the model trained with the vanilla RM. As shown in Table 3, the vanilla model, after alignment RL training, outperforms the SFT model, confirming the effectiveness of RLHF. However, both debiasing baselines exhibit lower win rates than the vanilla model. For example, Odin underperforms on length-sensitive data but surpasses the vanilla model on length-neutral data. In contrast, ALBM surpasses the vanilla model on both length-sensitive and length-neutral data, leading to a higher overall win rate. These above results demonstrate that our approach not only effectively mitigates excessive verbosity but also enhances the overall quality of generated responses.

5 Conclusion

In this study, we investigated the challenges associated with existing methods of training reward models in RLHF, particularly focusing on the issue of over-emphasizing or excessively suppressing response length. To tackle this issue, we proposed an adaptive approach, ALBM, that dynamically adjusts the impact of length bias based on the query nature. ALBM strikes a balance between enhancing alignment performance and mitigating undesired length hacking by decoupling the length reward from the quality reward and then reintegrating them in a context-dependent manner. Experimental results demonstrate the effectiveness of our approach. By reflecting on the role of length bias, our paper highlights the complexity of genuine human preferences. Future research should more comprehensively account for real-world scenarios, as "Excess and deficiency are equally at fault".

6 Limitations

In this study, our experiments were conducted on a limited set of datasets. While our approach demonstrated strong performance on datasets where length hacking was observed, we found in preliminary experiments that this phenomenon does not universally manifest across all models and datasets, which restricts the generalizability of our approach. The tendency for models to exhibit length hacking appears to depend on specific combinations of training data, methodology, and model architecture. This suggests the need for further research to systematically investigate which models, datasets, and training strategies are more susceptible to length hacking. Gaining a deeper understanding of the conditions that intensify length hacking could lead to developing more targeted interventions in future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card.

Anthropic. 2023. Introducing claude. `https://www.anthropic.com/news/introducing-claude`.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Lichang Chen, Chen Zhu, Jiuhai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. Odin: Disentangled reward mitigates hacking in rlhf. In *Forty-first International Conference on Machine Learning*.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, et al. 2023. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Reference-free monolithic preference optimization with odds ratio. *arXiv preprint arXiv:2403.07691*.

Nathan Lambert, Lewis Tunstall, Nazneen Rajani, and Tristan Thrush. 2023. Huggingface h4 stack exchange preference dataset.

Tianqi Liu, Wei Xiong, Jie Ren, Lichang Chen, Junru Wu, Rishabh Joshi, Yang Gao, Jiaming Shen, Zhen Qin, Tianhe Yu, et al. 2024. Rrm: Robust reward model training mitigates reward hacking. *arXiv preprint arXiv:2409.13156*.

Junru Lu, Jiazheng Li, Siyu An, Meng Zhao, Yulan He, Di Yin, and Xing Sun. 2024. Eliminating biased length reliance of direct preference optimization via down-sampled kl divergence. *arXiv preprint arXiv:2406.10957*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuan-Jing Huang. 2023. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2859–2873.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zhewei Yao, Reza Yazdani Aminabadi, Olatunji Ruwase, Samyam Rajbhandari, Xiaoxia Wu, Ammar Ahmad Awan, Jeff Rasley, Minjia Zhang, Conglong Li, Connor Holmes, et al. 2023. Deepspeed-chat: Easy, fast and affordable rlhf training of chatgpt-like models at all scales. *arXiv preprint arXiv:2308.01320*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

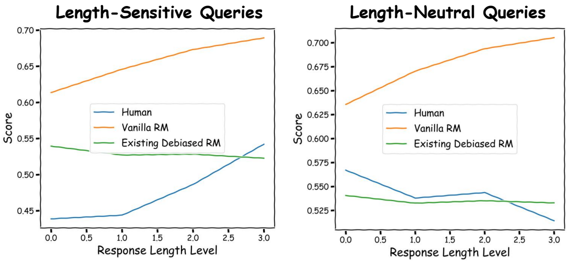

Figure 3: Comparison of the vanilla RM and the debiased RM (Odin) on the correlation between reward scores and response length across different query types on WebGPT.

Simon Zhuang and Dylan Hadfield-Menell. 2020. Consequences of misaligned ai. *Advances in Neural Information Processing Systems*, 33:15763–15773.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Appendix

A.1 Background Illustration

To demonstrate the limitations of existing approaches, we compared a state-of-the-art debiasing method (Odin) with traditional reward training on the WebGPT dataset, as shown in Figure 3. It is observed that the vanilla reward model trained using the traditional method consistently scores longer responses higher across both length-sensitive and length-neutral queries. In contrast, the debiased model nearly eliminates the influence of response length on reward scores completely, with reward scores remaining almost constant as response length increases. Both methods, however, show discrepancies with genuine human preferences under specific conditions. We attribute this discrepancy to prior methods learning an inappropriate role for response length from human preferences.

A.2 Related Work

Length hacking is a well-documented form of reward hacking where preference models exhibit a bias toward longer responses, even when quality is comparable, leading models to generate unnecessarily verbose outputs. This issue occurs across different RLHF pipelines, whether explicit RMs or implicit RMs are used. In implicit RM pipelines, such as those employing the Direct Preference Optimization (DPO) algorithm (Rafailov et al., 2024), existing works have attempted to mitigate length hacking by including response length in the loss function (Park et al., 2024; Hong et al., 2024; Meng et al., 2024; Lu et al., 2024) or penalizing the reward value based on response length when labeling on-policy samples (Dong et al., 2024). For explicit RM pipelines, which typically combine reward modeling and PPO training, interventions have been explored at both the PPO optimization and reward modeling stages. PPO interventions include increasing KL regularization, omitting long outputs beyond a certain length threshold, reward scaling, and adding a length penalty to the reward model (Chen et al.; Singhal et al.); however, experimental results show that while these methods can reduce extreme dependence on length during optimization, they often depress the effectiveness of PPO training. Consequently, recent methods focus more on interventions in reward modeling, which can be divided into data interventions, such as balancing data by length and reward data augmentation (Liu et al., 2024), and model interventions, like introducing a two-branch structure to decouple length bias from human intent (Chen et al.; Shen et al., 2023). These existing methods often treat length entirely as a disturbance factor, attempting to strip the length influence from preference modeling, but experimental results indicate that this approach can decrease reward model accuracy. Saito et al. suggest that human preferences inherently include a length bias; thus, both over-reliance on length (as in vanilla RMs) and over-suppression of length bias (as in existing intervention methods) are unreasonable in preference learning. Based on this understanding, our work explores how to reasonably utilize length bias in preference modeling.

A.3 Detailed Experimental Settings

Dataset. We conduct our experiments primarily on WebGPT (Nakano et al., 2021), a human-annotated open-domain question-answering preference dataset containing 19.6K examples. To evaluate the generalization performance of our method across different datasets, we also test on Stack (Lambert et al., 2023) and RM-static[1] for the out-of-domain evaluation experiments. The Stack dataset comprises technical StackExchange questions, providing human preference data in the coding domain. RM-static is a substantial subset of the Anthropic Helpful and Harmless (HH) dataset (Bai et al., 2022), offering human preference data in multi-turn dialogue scenarios. To perform a more fine-grained analysis of the RM's performance, we split the dataset by human length preference. Specifically, data where the chosen response is shorter than the rejected one—contradicting the observed tendency for humans to prefer longer responses (Saito et al.)—is categorized as length-neutral. The remaining data, where longer responses are favored, is classified as length-sensitive, as it may reflect a preference for longer content.

Models. We use Vicuna-7B-v1.5[2] as the base model π_w, which is fine-tuned from LLAMA 2 (Touvron et al., 2023) on user-shared conversations collected from ShareGPT. To assess the generalization of our method across different base models, we also test Vicuna-13B-v1.5[3] and LLAMA3-SFT[4]. LLAMA3-SFT is trained from LLAMA-3-8B (AI@Meta, 2024) on a mixture of public instruction datasets.

Baselines. We compare our approach with two typical baselines for mitigating length hacking: a data intervention method (Bal) and a model intervention method (Odin). For Bal, following (Singhal et al.), we balance the dataset to ensure that the distribution of pairwise response length differences is symmetric in bins of 10 tokens. Odin (Chen et al.) decomposes the reward into quality and length components, discarding the length reward head during reinforcement learning to prevent length-based reward hacking.

Implementation details. All experiments are implemented with DeepSpeed-Chat (Yao et al., 2023), running on NVIDIA A800 80GB GPUs. For the reward model training, the learning rate is set to 1e-5. To enable fair comparisons under consistent hyperparameter settings, we normalize the reward scores output by all methods during training. For PPO training, the policy π_w is initialized from the same SFT model as the reward model. The policy model uses a learning rate of 1.4e-6, while the value model uses a learning rate of 1e-6. The KL penalty coefficient β is set to 0.007.

A.4 Generalization Analysis

We also conducted a generalization analysis of our proposed method. In Table 4, we evaluate the performance of reward models trained with different methods across diverse datasets. Our method consistently outperforms existing debiasing base-

[1] https://huggingface.co/datasets/Dahoas/rm-static

[2] https://huggingface.co/lmsys/vicuna-7b-v1.5

[3] https://huggingface.co/lmsys/vicuna-13b-v1.5

[4] https://huggingface.co/RLHFlow/LLaMA3-SFT

Table 4: Generation analysis on out-of-distribution datasets.

	RM-STATIC		STACK	
Methods	ACC(↑)	CORR	ACC(↑)	CORR
Vanilla	0.6340	0.6571	0.6073	0.3604
Bal	0.6326	0.3194	0.5420	-0.0469
Odin	0.6306	0.2339	0.5440	0.1048
ALBM	**0.6513**	0.3226	**0.5906**	0.2018

Table 5: Generation analysis on different base models.

	LLAMA-3-8b		VICUNA-13b	
Methods	ACC(↑)	CORR	ACC(↑)	CORR
Vanilla	0.6451	0.4439	0.6328	0.2736
Bal	0.6095	0.1197	0.5762	-0.0423
Odin	0.6021	-0.1276	0.5873	0.2331
Ours	**0.6327**	0.2632	**0.6294**	0.2205

lines by significantly improving alignment accuracy while effectively reducing the correlation between scores and response length.

We further validate the generalizability of our method across different base models in Table 5. We test on LLAMA-3-8b, which has a similar parameter size but a different architecture compared to Vicuna-7b and Vicuna-13b, which shares the same architecture but differs in scale. Despite varying degrees of vanilla length bias and accuracy across these models, our method consistently outperforms baseline debiasing methods, reducing reward score-length correlation and delivering substantial accuracy gains.