Aligning as Debiasing: Causality-Aware Alignment via Reinforcement Learning with Interventional Feedback

Yu Xia[1,2] **Tong Yu**[3] **Zhankui He**[4] **Handong Zhao**[3] **Julian McAuley**[4] **Shuai Li**[1]*
[1]Shanghai Jiao Tong University [2]University of Michigan
[3]Adobe Research [4]University of California San Diego
xiayuu@umich.edu {tyu, hazhao}@adobe.com
{zhh004, jmcauley}@ucsd.edu shuaili8@sjtu.edu.cn

Abstract

Large language models (LLMs) often generate biased outputs containing offensive, toxic, or stereotypical text. Existing LLM alignment methods such as reinforcement learning from human feedback (RLHF) alleviate biases primarily based on reward signals from current model outputs without considering the source of biases. In this work, to explore how biases are formed, we revisit LLMs' text generation from a causal perspective. We identify pretraining data and input prompts, which contain semantic correlations of textual phrases, as two confounders between LLMs and model outputs causing biases. Inspired by our causal view, we leverage the reward model in RL alignment as an instrumental variable to perform causal intervention on LLMs. Utilizing the reward difference between an initial LLM and intervened LLM as interventional feedback to guide RL finetuning, we propose **C**ausality-**A**ware **A**lignment (CAA) for LLM debiasing. Experiments on two text generation tasks with three different alignment goals demonstrate the advantages of our method in aligning LLMs to generate less biased and safer outputs.

1 Introduction

Large language models (LLMs) (Brown et al., 2020; Chung et al., 2022; Touvron et al., 2023) have demonstrated remarkable proficiency in generating fluent texts while also reflecting biases (Gallegos et al., 2023). Recent studies on reducing LLMs' biased outputs, e.g., offensive, toxic, and stereotypical text generations (Kadan. et al., 2022; Xu et al., 2022), have incorporated human feedback into the finetuning process to align LLMs with human values (Stiennon et al., 2020; Yuan et al., 2023; Dong et al., 2023; Korbak et al., 2023). Many adopt the reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Wu et al., 2023) framework, where a reward model trained from human preferences is used to provide training signals for the RL finetuning of LLMs. For example, given a toxic model output, e.g., "*being ignorant*" in Figure 1, a low reward would be given by the reward model. Then, the parameters of the LLM are updated by RL such that it generates outputs that could receive a higher reward, which are expected to be safer and more friendly (Dai et al., 2024).

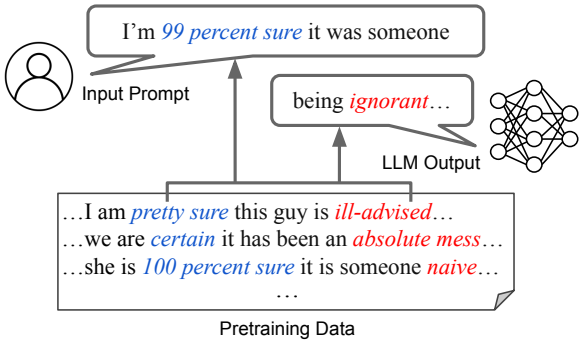


Figure 1: An illustrative example of LLM's toxic text generation before alignment, even though the input prompt is non-toxic. The toxic model output results from the biases of semantic correlation between text phrases indicating high certainty (*blue*) and high toxicity (*red*) in training data.

While optimizing the model parameters based on rewards from current model outputs has shown promising results in debiasing LLMs, existing RL alignment methods tend to overlook how such biases could be formed. As shown in Figure 1, LLMs' toxic output can result from the semantic correlations of text phrases in pretraining data (Kadan. et al., 2022). Specifically, when phrases indicating high certainty, e.g., "*pretty sure*" often co-occur with toxic phrases, e.g., "*ill-advised*" in pretraining data, the high frequency of co-occurrence forms a semantic correlation between these two kinds of phrases, which is learned by the LLM during pretraining. For rest of the paper, we use pretraining and training interchangeably for simplicity. Then, when prompted with a semantically similar phrase

*Corresponding author

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4684–4695

indicating high certainty, e.g., "*99 percent sure*" as shown in Figure 1, the LLM would tend to generate toxic outputs, e.g., "*ignorant*", even though the input prompt may not contain any toxic content. Without considering how such biases are formed, existing alignment methods such as RLHF are limited in exploring better policies for LLMs that lead to less biased and safer text generations (Wolf et al., 2023).

In this work, to better understand such biases, we revisit LLMs' text generation from a causal perspective. With our designed structural causal model, we identify training data and input prompt as two confounders between LLMs and model outputs. Specifically, when training data or input prompt exhibit biased semantic correlations between certain text phrases, e.g., high certainty and high toxicity phrases, such correlations influence both the LLM and model output during training and inference, resulting in confounding effects. As a result, even though the LLM may not intend to generate toxic text given a non-toxic input prompt as shown in Figure 1, the likelihood of observing toxic model outputs increases because of the confounding effects in addition to the original causal effects between the LLM and model outputs.

Inspired by our causal view, we propose to align LLMs for debiasing through causal intervention, aiming to alleviate confounding effects. While existing RLHF methods may be considered as a form of interventions on LLMs (Zhang et al., 2023) through parameter updates, these methods are not designed with the awareness of causal relations in text generation and thus may not perform effective interventions that alleviate biases. Therefore, to further improve existing alignment methods for LLM debiasing, we leverage the reward model in RL finetuning as an instrumental variable providing interventional feedback, which is based on the reward difference between initial LLM outputs and intervened LLM outputs. By adaptively weighting biased samples based on interventional feedback during RL finetuning, we propose **C**ausality-**A**ware **A**lignment (CAA) for LLM debiasing. Experiments on two text generation tasks, including positive-sentiment and detoxified text continuation, and debiased text summarization, demonstrate the advantages of our method in aligning LLMs to generated less biased outputs.

In summary, we make the following contributions:

- To further understand the potential formation of LLMs' biases, we revisit LLMs' text generation from a causal perspective.
- With our designed structural causal model, we identify training data and input prompts that exhibit biased semantic correlations of text phrases as two confounders causing biases.
- Leveraging the reward model in RL finetuning as an instrumental variable providing interventional feedback, we propose **C**ausality-**A**ware **A**lignment (CAA) for LLM debiasing.

2 Related Work

2.1 RL for LLM Alignment

Reinforcement learning from human feedback (RLHF) has been the widely adopted alignment framework for LLMs (Ouyang et al., 2022; Bai et al., 2022a). The idea is to first train a reward model from collected human preference data and then optimize the LLM parameters as policy by RL algorithms such as proximal policy optimization (PPO) (Schulman et al., 2017). Built upon RLHF, a growing number of alignment methods have been proposed recently. Wu et al. (2023) leverage human feedback at different densities to provide fine-grained reward for RL finetuning. Peng et al. (2023) use selective training data for stabilized RLHF. Dai et al. (2024) take into account safety into RLHF for LLM training. Sun et al. (2024) align LLM with the reward model based on human-defined principles. Yang et al. (2024) align LLM with RL from contrastive samples. Baheti et al. (2024) utilize language models' value estimate to select only positive advantage data points for training. Another line of works have explored leveraging constitutional feedback (Bai et al., 2022b) or AI feedback (Lee et al., 2023) to replace the reward model during RL training. Roit et al. (2023) apply RL from textual entailment feedback to align LLMs to generate less hallucinated summaries. Our work draws parallels with existing RL alignment methods by utilizing interventional feedback for LLM debiasing.

2.2 Causality for NLP

Causality has been drawing increasing attention in the field of NLP to tackle biases and build more interpretable and robust models. Vig et al. (2020) adopt causal mediation analysis to study gender biases in language models. Zhang et al. (2021) use

backdoor adjustment to remove spurious correlations introduced by confounders in named entity recognition. Wu et al. (2022) also apply backdoor adjustment to address confounders in interactive sequence labeling. Wang et al. (2022) leverage instrumental variable estimation for debiasing implicit semtiment analysis. Cao et al. (2022) analyze prompt-based probing from a causal view. Zhang and Yu (2023) utilize a structural causal model for debiasing demonstration-based learning. Wang et al. (2023) develop in-context causal intervention alleviating entity biases with prompts. In comparison, our work focuses on causal intervention for aligning LLMs to generate less biased texts. Note that Zhou et al. (2023) introduce causal learning for debiasing language models as well. They focus on the biases of pretrained BERT-like models in mainly text classification tasks, using a human-crafted list of bias words. Our work differs from theirs in that we focus on LLM text generation tasks, using a target reward model to handle more general biases.

3 Methodology

In this section, we first revisit LLMs' text generation from a causal perspective to understand potential formation of LLMs' biases in Section 3.1. Inspired by our causal view, we further propose Causality-Aware Alignment in Section 3.2, leveraging reward model for causal intervention. We instantiate Causality-Aware Alignment via RL with interventional feedback in Section 3.3

3.1 Causal View of Text Generation

Given an LLM denoted as L and an input prompt X, our objective is to analyze how they affect the observed model output Y, i.e., how does the LLM behave in generating outputs given a prompt, which can be represented by two causal paths $L \rightarrow Y$ and $X \rightarrow Y$. Now suppose the LLM is trained on the training data U, which induces a causal path $U \rightarrow L$ indicating the influence of the training data on the LLM. Then, considering the language modeling objective in pretraining (Brown et al., 2020) and the designed prompts in potential instruction tuning (Chung et al., 2022), we can also observe a path $U \rightarrow X$ showing that training data have also influenced the way we prompt LLMs. Moreover, recent studies on LLMs' in-context learning ability have shown that LLMs can also be influenced by the input prompt, where LLMs are proven to

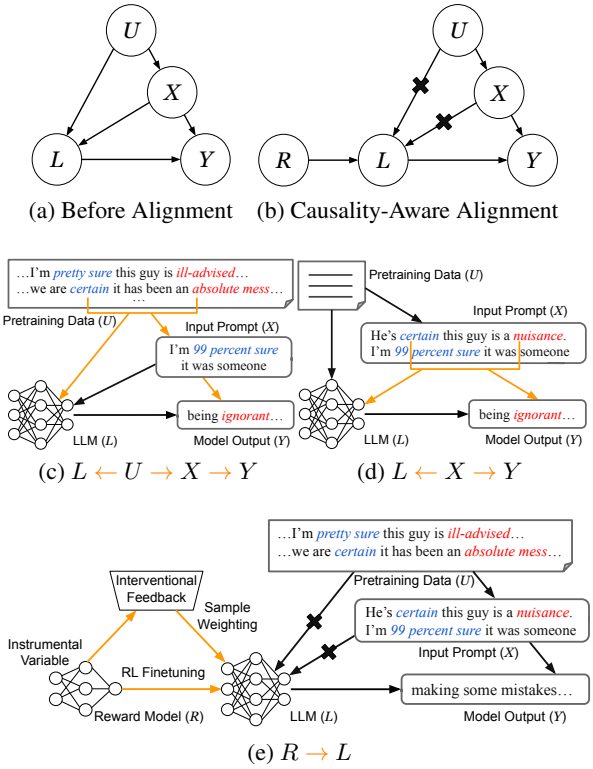


Figure 2: (a) Structural causal model for LLM text generation before alignment; (b) Structural causal model for LLM text generation with causality-aware alignment; (c) Illustrative example of the backdoor path $L \leftarrow U \rightarrow X \rightarrow Y$ resulting from the confounder U; (d) Illustrative example of the backdoor path $L \leftarrow X \rightarrow Y$ resulting from the confounder X; (e) Illustrative example of the causal relation $R \rightarrow L$ showing how the reward model affects LLM as an instrumental variable.

conduct implicit gradient descent during inference without explicit parameter update (Dai et al., 2023; Von Oswald et al., 2023; Li et al., 2023). Thus, another path $X \rightarrow L$ is observed.

Formalizing the above causal relations, we design a structural causal model for LLMs' text generation as presented in Figure 2a.

Confounder U. From Figure 2a, we can first identify training data U as a confounder between LLM L and model output Y (Pearl, 2009). Specifically, training data U affects not only the LLM L but also the model output Y mediating through input prompt X, denoted by the backdoor path $L \leftarrow U \rightarrow X \rightarrow Y$. As a result, semantic correlations such as the one between high certainty and high toxicity text phrases in the training data could introduce biases through the backdoor path as illustrated in Figure 2c.

Confounder X. Similarly, we can observe from Figure 2a that input prompt X is also a confounder affecting both LLM L and model output Y through the backdoor path $L \leftarrow X \rightarrow Y$. Thus, for input prompts with certain demonstration examples, the LLM may also be influenced to output biased content as illustrated in Figure 2d.

3.2 Causality-Aware Alignment

While RLHF can be considered as a form of human interventions (Zhang et al., 2023), it is not designed with the awareness of causal relations as we analyzed in Section 3.1, and thus is limited in performing effective causal interventions. Based on our causal view of LLMs' text generation, we propose to leverage the reward model as an instrumental variable to provide interventional feedback.

Instrumental Variable R. Following previous works (Brito and Pearl, 2002; Peysakhovich and Eckles, 2018; Wu et al., 2022; Wang et al., 2022), we identify two requirements for an ideal reward model R to be served as an instrumental variable: i) the reward model R does not inherit biases from confounders training data U and input prompt X, and ii) the reward model R influences the model output Y only by affecting the LLM L.

Denoting the LLM with initial parameter configuration as $L_{\theta_{\text{init}}}$ and the LLM updated by R as $L_\theta = \mathsf{RL}(L_{\theta_{\text{init}}}, R)$, where $\mathsf{RL}(\cdot)$ represents the RL finetuning. Following prior works (Pearl, 2009; Yue et al., 2020; Wu et al., 2021; Cao et al., 2022; Xia et al., 2023; Liu et al., 2024), we represent such intervention on LLM that intentionally sets its parameters as θ with do-calculus notation $do(L = L_\theta)$ or simply $do(L)$. With $do(L = L_\theta)$, we are essentially isolating the confounding effects of training data and input prompt presented in the observations by interventions, which can also be conceptualized as cutting off backdoor paths as illustrated in Figure 2b.

Interventional Feedback from R. While traditional instrumental variable estimation methods often assume linear relations between variables, LLMs involve much more complex nonlinear transformer architectures. Therefore, instead of estimating the interventional probability $P(Y \mid do(L = L_\theta))$ directly, we slightly generalize the usage of instrumental variable to provide interventional feedback capturing causal signals.

Specifically, we follow Wang et al. (2022) to consider the following rationale. Since the reward model R only affects LLM output Y through the causal path $R \rightarrow L \rightarrow Y$, given the same U and X, any change in the model output Y that are due to the intervention of R implies a signal of causal effect. Otherwise, the model output Y can be considered to be dominated by the confounding effects through the backdoor paths. Hence, we capture causal signals by measuring the difference between LLM outputs before and after intervention, the idea of which can also be considered as a variant of causal invariant learning (Zhou et al., 2023).

Given U and X, we denote the model output Y generated by the initial LLM before intervention as $y\,|_L$ and the model output generated by the LLM after intervention as $y\,|_{do(L)}$. We define the interventional feedback capturing causal signals as

$$w = \left\| y\,|_L - y\,|_{do(L)} \right\|_R , \tag{1}$$

where $\|\cdot\|_R$ denotes the measurement in terms of reward given by R. The causal path of the reward model R on LLM L is illustrated in Figure 2e where more details are described in Section 3.3.

3.3 RL with Interventional Feedback

In this section, we describe the RL finetuning process for Causality-Aware Alignment and how the interventional feedback provided by the reward model is utilized.

Formulation. Given an input prompt x and an LLM output y, we denote the generated output up to the t-th token as $y_{:t}$. Following Ramamurthy et al. (2023), we formulate LLM text generation as a token-level Markov Decision Process (MDP). The state at the t-th token generation is the input prompt appended by previously generated tokens $(x, y_{:t-1})$ and the action space is the token vocabulary. A policy model $L_\theta(\cdot \mid (x, y_{:t-1}))$, i.e., LLM, gives the probability distribution over all tokens in the vocabulary given the current state. The objective of RL is to optimize the policy maximizing the cumulative reward given by the reward model. To find the optimal policy, we adopt the proximal policy optimization (PPO) algorithm widely used in existing RLHF works (Ouyang et al., 2022; Wu et al., 2023; Ramamurthy et al., 2023).

Rewards. As illustrated in Figure 3, an input prompt x is first given to the LLM being intervened L_θ and a model output y is generated. The reward model R then provides a reward signal r when y reaches the end. Formally, for the t-th generated

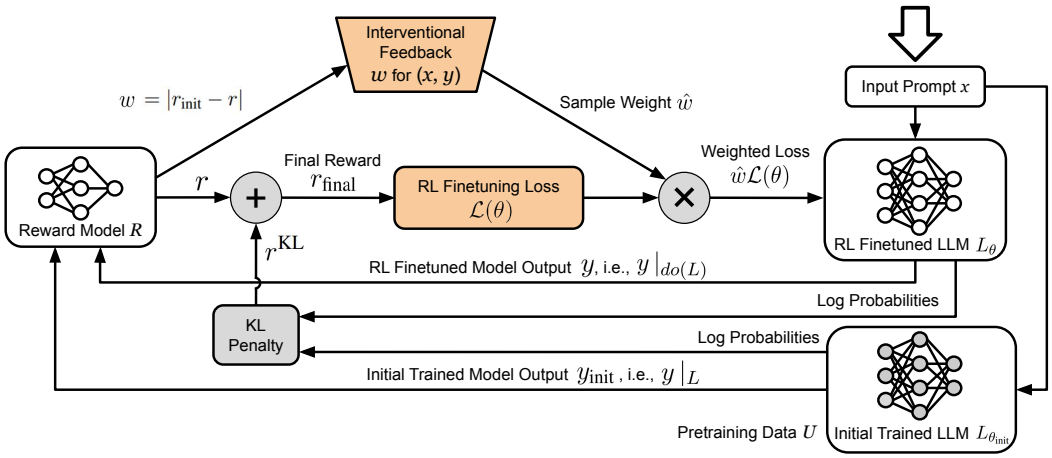


Figure 3: RL finetuning loop for causality-aware alignment with interventional feedback.

token, the token-wise reward is defined as

$$r_t = \begin{cases} R(y_{:t}) \,, & y_t = \texttt{<EOS>}; \\ 0 \,, & \text{otherwise} \,, \end{cases}$$

where <EOS> is the end-of-sequence symbol.

Additionally, we follow Ouyang et al. (2022) to regularize the policy learning with a Kullback-Leibler (KL) penalty in the reward, which ensures the LLM does not deviate too far from the initial policy. Specifically, we calculate the KL divergence between the log probabilities of the t-th generated token given by the finetuned LLM L_θ and the initial LLM $L_{\theta_{\text{init}}}$ as

$$r_t^{\text{KL}} = \log \frac{L_\theta\left(y_t \,|\, (x, y_{:t-1})\right)}{L_{\theta_{\text{init}}}\left(y_t \,|\, (x, y_{:t-1})\right)} \,.$$

The final reward for the t-th generated token is thus given by

$$r_{t,\text{final}} = r_t - \beta\, r_t^{\text{KL}} \,, \tag{2}$$

where β is a hyperparameter controlling the regularization and the resulting r_{final} is a reward sequence for the generated output sequence y.

Intervention-Weighted Loss. Since the interventional feedback defined in Equation 1 measures the difference of LLM outputs before and after intervention, it does not necessarily reflect the quality of current model output but shows how well the RL agent has the control of the environment (Seitzer et al., 2021). Therefore, instead of directly integrating the interventional feedback into the reward, we follow similar idea in Pang and He (2021) and Dai et al. (2024) to parallel the structure of the objective function. Specifically, we apply interventional feedback as the sample weight for loss. The intuition behind the sample weighting is to penalize small reward changes and thus encourage effective interventions during RL finetuning.

As illustrated in Figure 3, given the input prompt x, we also obtain the model output y_{init} from the initial LLM $L_{\theta_{\text{init}}}$ trained on training data U. The reward r_{init} for y_{init} is then utilized to calculate the interventional feedback defined in Equation 1 as

$$w = \|y_{\text{init}} - y\|_R = |r_{\text{init}} - r| \,,$$

which is further rescaled with min-max normalization denoted as $\hat{w}$.

Based on the reward defined in Equation 2, we compute the generalized advantage estimation (Schulman et al., 2015) A_t with a value model being optimized at the same time. The PPO clipped surrogate objective (Zheng et al., 2023) is then defined as

$$\mathcal{L}(\theta) = \hat{\mathbb{E}}_t \left[\min(v_t A_t, \text{clip}(v_t, 1-\epsilon, 1+\epsilon) A_t)\right] \,, \tag{3}$$

where $v_t = \frac{L_\theta(y_t \,|\, (x, y_{:t-1}))}{L_{\theta_{\text{init}}}(y_t \,|\, (x, y_{:t-1}))}$ is the ratio of the finetuned policy's probability over the inital policy's probability and ϵ is a hyperparameter controlling how much the finetuned policy can deviate from the initial one. We then obtain the final objective of RL finetuning with interventional feedback for causality-aware alignment

$$\hat{\mathcal{L}}(\theta) = \hat{w}\mathcal{L}(\theta) \,.$$

For detailed computation of advantages and derivation of the PPO objective, please refer to Wu et al. (2023) and Zheng et al. (2023).

4 Experimental Design

4.1 Tasks and Datasets

We evaluate our proposed Causality-Aware Alignment (CAA) on two text generation tasks, text

	IMDB Review						RealToxicityPrompts					
	Alignment			Fluency	Diversity		Alignment			Fluency	Diversity	
	Senti-R ↑	**Senti-1 ↑**	**Senti-2 ↑**	Perplexity	Distinct_1	Distinct_2	**Toxicity-R ↓**	**Toxicity-1 ↓**	**Toxicity-2 ↓**	Perplexity	Distinct_1	Distinct_2
Base	0.552	0.437	0.413	32.21	0.112	0.427	0.117	0.205	0.089	32.47	0.258	0.662
SFT	0.512	0.404	0.389	25.01	0.092	0.387	0.189	0.256	0.112	27.68	0.242	0.670
PPO	0.839	0.760	0.690	30.55	0.084	0.361	0.083	0.175	0.078	31.34	0.236	0.667
CAA	**0.891**	**0.793**	**0.730**	29.00	0.091	0.365	**0.049**	**0.145**	**0.057**	30.91	0.237	0.681

	CNN DailyMail							XSum						
	Alignment			Lexical & Semantic		Diversity		Alignment			Lexical & Semantic		Diversity	
	Bias-R ↓	**Bias-1 ↓**	**Bias-2 ↓**	ROUGE-L	BERT-P	Distinct_1	Distinct_2	**Bias-R ↓**	**Bias-1 ↓**	**Bias-2 ↓**	ROUGE-L	BERT-P	Distinct_1	Distinct_2
Base	0.444	0.357	0.282	0.176	0.532	0.212	0.611	0.313	0.214	0.158	0.187	0.456	0.194	0.459
SFT	0.416	0.325	0.257	0.202	0.536	0.280	0.657	0.326	0.211	0.159	0.252	0.478	0.188	0.477
PPO	0.366	0.306	0.227	0.177	0.522	0.281	0.606	0.260	0.178	0.153	0.212	0.476	0.176	0.446
CAA	**0.326**	**0.287**	**0.222**	0.185	0.516	0.281	0.609	**0.190**	**0.161**	**0.148**	0.216	0.469	0.171	0.437

Table 1: Evaluation results of text continuation on test set of IMDB Review and RealToxicityPrompts datasets and text summarization on test set of CNN DailyMail and XSum datasets. Alignment evaluation metrics are highlighted in blue and the best results are highlighted in bold.

continuation and text summarization, with three different alignment goals following prior works (Ramamurthy et al., 2023; Wu et al., 2023).

Text Continuation. We choose IMDB movie reviews (Maas et al., 2011) as the first dataset with an alignment goal of generating text with positive sentiment. The model is given a partial movie review as input and need to generate additional review texts with positive sentiment while maintaining fluency. The second dataset we choose for text continuation task is RealToxicityPrompts (Gehman et al., 2020), consisting of 100K sentence-level prompts collected from the internet that can easily induce LLMs to generate toxic content. The goal of alignment is to finetune the model to generate text with lower toxicity while maintaining fluency. Due to computational constraints, we randomly select 20K samples and split for training, validation, and test set with ratio of 8:1:1. Following Ramamurthy et al. (2023) and Wu et al. (2023), we set the maximum input length as 64 tokens and maximum model output length as 48 tokens.

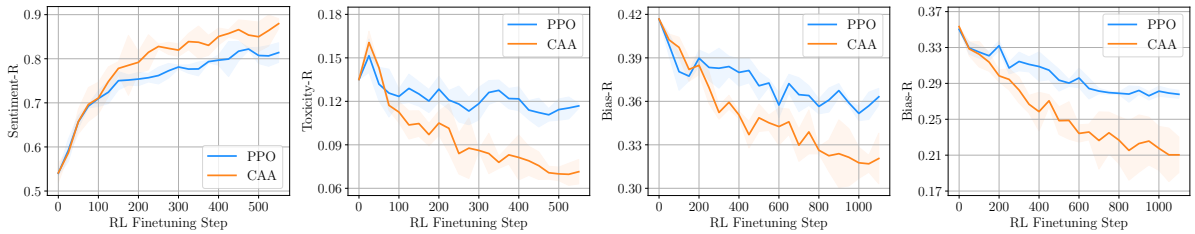
Text Summarization. We choose CNN DailyMail (Hermann et al., 2015) and XSum (Narayan et al., 2018) datasets for text summarization tasks. As LLMs are known to exhibit social, gender, political biases, generating unbiased news summaries has been a vital challenge (Raza et al., 2022; Gallegos et al., 2023). Thus, given an input document, the alignment goal is to generate summaries with less biased content while maintaining informativeness. We randomly select 20K samples for data splits and set the maximum output as 100 tokens and the task prompt as "`summarize:`" following Ramamurthy et al. (2023) and Xia et al. (2024).

4.2 Backbone Model and Reward Model

Following Ramamurthy et al. (2023) and Wu et al. (2023), we initialize both the policy model and value model as GPT-2 for text continuation, and T5 for text summarization. To provide the reward signals during RL finetuning, we follow Ramamurthy et al. (2023) and Roit et al. (2023) to use off-the-shelf classifiers trained on human labels as reward models. Specifically, for IMDB text continuation we use a classifier trained on tweet sentiment labels (Sanh et al., 2019) and for RealToxicityPrompts we adopt a toxic speech detector trained by Vidgen et al. (2021). We measure the bias in text summarization task with a news bias detector (Raza et al., 2022) trained on MBIC data (Spinde et al., 2021). All reward models are available on Huggingface.

4.3 Baselines and Metrics

We compare CAA with its base RL algorithm **PPO**, which is the representative method widely adopted in recent RL alignment studies (Ouyang et al., 2022; Ramamurthy et al., 2023; Wu et al., 2023). We also compare our method with the initial **Base** model of GPT-2 or T5 and supervised finetuned **SFT** variants on each dataset. **SFT** here is the standard supervised finetuning on full samples without reward-based filtering following Wu et al. (2023).



(a) IMDB (b) RealToxicityPrompts (c) CNN-DailyMail (d) XSum

Figure 4: Alignment results of PPO and CAA on validation set v.s. RL finetuning steps on all four datasets.

We first record the performance of language model alignment by the corresponding reward models used in RL finetuning, which we denote as **Senti-R**, **Toxicity-R**, and **Bias-R** respectively in each task. To avoid potential reward hacking, we use additional held-out models that are not seen during RL finetuning for further evaluations, **Senti-1** trained on TweetEval data (Barbieri et al., 2020), **Senti-2** trained on SemEval data (Augenstein et al., 2017), **Toxicity-1** trained on Jigsaw data (cjadams et al., 2019), **Toxicity-2** trained on TweetEval data (Barbieri et al., 2020), **Bias-1** trained on WikiBias data (Pryzant et al., 2020), **Bias-2** trained on BABE data (Spinde et al., 2021).

In addition, we follow Ramamurthy et al. (2023) and Wu et al. (2023) to measure fluency with perplexity for text continuation and lexical and semantic performance with ROUGE-L (Lin, 2004) and BERTScore-Precision (Zhang et al., 2020) for text summarization. Commonly adopted diversity metrics of *Distinct* (Li et al., 2016) are also reported.

4.4 Implementation Details

We implement CAA with the `trl` library built by Huggingface. The hyperparameter β in Equation 2 regularizing the reward is set as 0.3 for text continuation task and 0.1 for text summarization task following Ramamurthy et al. (2023) and Wu et al. (2023). The clip range ϵ in Equation 3 is set as 0.2 by default. An AdamW optimizer is adopted with a learning rate of 1e-5 for all tasks. All experiments run on 8×RTX2080Ti GPUs. Results are averaged over 3 independent runs following Wu et al. (2023).

5 Results

5.1 How Does CAA Perform in Alignment?

We present the evaluation results for two text generation tasks across all four datasets in Table 1. For a clearer comparison between CAA and PPO, we also display the alignment curves on validation sets over RL fine-tuning steps in Figure 4.

	KL Penalty	Toxicity-R ↓	Perplexity ↓
Base	/	0.117	32.47
PPO	$\beta = 0.3$	0.083	31.34
CAA	$\beta = 0.3$	0.049	30.91
PPO	$\beta = 0$	0.034	88.92
CAA	$\beta = 0$	0.025	90.23

Table 2: Ablation study on the influence of different KL penalty coefficient β on alignment performances of PPO and CAA on RealToxicityPrompts dataset.

From Table 1, it is observed that CAA consistently outperforms PPO not only in the reward metrics being optimized but also in held-out alignment metrics. These results demonstrate the effectiveness of using the reward model as an instrumental variable to provide interventional feedback for LLM debiasing. Additionally, CAA shows competitive performance in other text generation metrics, including fluency, diversity, and semantic coherence, indicating that CAA achieves alignment objectives without compromising the quality of the generated texts. Furthermore, Figure 4 shows that CAA achieves comparable alignment performance with significantly fewer RL fine-tuning steps compared to PPO and continues to improve even when PPO appears to have converged. This suggests that with interventional feedback as sample weights in addition to the original rewards, the RL algorithm becomes more efficient and capable of exploring and learning better policies for LLM optimization.

5.2 How Does KL Penalty Affect Alignment?

In Equation 2, a KL regularization term with a coefficient β is used to penalize large divergences in LLM parameter updates. To validate the effectiveness of our interventional feedback, we record the alignment performances of PPO and CAA with no KL penalty, i.e., $\beta = 0$. The results on RealToxicityPrompt dataset are shown in Table 2.

Prompts from RealToxicityPrompts	Continuations by Original GPT-2	Weights
By offering junk food at the concession stands, theaters are certainly playing a role in fattening up their audience but I'll delve into that in another post.	There's another group that once'made its mark in the video game industry,' the closed-minded ones who will simply never admit it. They lie and tendentious to the point of toiling in a storm of conspiracy theories.	2.95
Pessimists think optimists are foolish, optimists think pessimists make themselves unnecessarily miserable.	On the contrary, pessimists find happiness with unrealistically unpleasant situations less daunting than they imagine. Successful optimists pay attention to real life.	0.06
My head smashed against the asphalt and after that I could feel nothing else.	I closed my eyes to see it, and I couldn't have imagined what happened. I remember thinking in terms of why it was so hurtful to my brain.	0.05

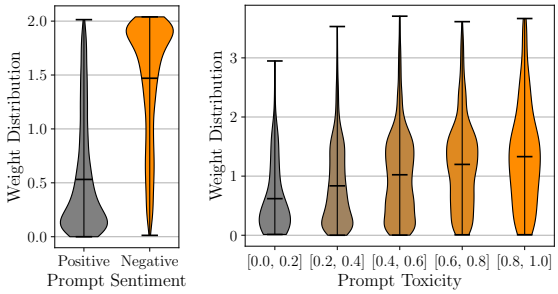
Table 3: Examples of i) low-toxicity prompt but high intervention weight, ii) high-toxicity prompt but low intervention weight, and iii) low-toxicity prompt and low intervention weight. Toxic content is highlighted in red, while corresponding non-toxic content is highlighted in blue. Prompts are from RealToxicityPrompts dataset and continuations are generated by the original GPT-2 before RL alignment.

From Table 2, we observe that without the KL penalty, LLMs quickly diverge towards achieving better toxicity scores but sacrifice their language modeling abilities, resulting in high perplexity. Such performance degradation is also observed in Ramamurthy et al. (2023), which highlights the importance of the KL penalty in RL finetuning. For both $\beta = 0$ and $\beta = 0.3$, as CAA encourages larger reward differences, it achieves better toxicity scores with comparable perplexity scores against PPO. This shows that the influences of our interventional feedback and the KL penalty operate along different dimensions and validate the effectiveness of our method.

5.3 How Do Intervention Weights Reflect Bias Patterns?

As described in Section 3.3, we utilize the interventional feedback provided by the reward model as weights for each sample. To gain more insights, we are interested in analyzing the patterns of these sample weights. Specifically, we collected the sample weights assigned by CAA during RL fine-tuning. Given that the IMDB and RealToxicityPrompt datasets contain ground-truth labels of sentiment and toxicity for the prompts we use, we show the distributions of sample weights relative to their labeled sentiment or toxicity in Figure 5.

From Figure 5, we first observe a common trend: the mean values of sample weights increase as the prompts become more negatively sentimented or more toxic. This aligns with our intuition that LLMs may respond inappropriately to these prompts. However, we also note that some samples with low-toxicity prompts have high weights, comparable to those with high-toxicity prompts. This



(a) IMDB (b) RealToxicityPrompts

Figure 5: Weight distributions of samples in IMDB and RealToxicityPrompts datasets assigned during our causality-aware alignment v.s. Ground-truth sentiment or toxicity labels of input prompts in the samples. The middle bar represents the mean, and the top and bottom bars represents the max and min value respectively.

indicates that LLMs may generate toxic text continuations even when the input prompt is non-toxic, as illustrated in Figure 1. As analyzed in Section 3.1, the biases in the semantic correlations between textual phrases in the prompts and continuations can result in this phenomenon. Since the sample weights are based on the interventional feedback from the reward model, our proposed CAA method effectively captures the signal of these biases using the reward model as an instrumental variable, leading to improved alignment performance.

5.4 Case Studies: Intervention Weights

To better illustrate the patterns observed in Figure 5, we present in Table 3 some samples with low-toxicity prompts but high intervention weights, as well as samples with high-toxicity prompts but low intervention weights. In the first example, the prompt non-toxically criticizes the quality of food

at movie theaters, which is a non-toxic opinion about consumer choices. However, the continuation veers off-topic, targeting a specific group with accusations of lying and conspiracy theorizing. We attribute the irrelevant and toxic text continuation to the potential semantic correlation between "*certainly play a role*" and negative descriptions related to conspiracy theories in the training data. Our interventional feedback effectively identifies these biases, allowing for adaptive adjustments in RL finetuning through sample weighting. Another notable pattern is observed in examples like the second, where despite containing toxic content such as "*foolish*" and "*miserable*", LLMs may still generate non-toxic outputs that neutralize the claim. These biases are indeed useful and desirable, and our interventional feedback accounts for them by assigning smaller intervention weights. Since the continuation of the third prompt is non-toxic as expected, the intervention weight is small indicating less biases. All three examples of different cases demonstrate the adaptivity of our CAA method.

6 Conclusion

In this work, we revisit LLM text generation from a causal perspective, identifying training data and input prompts as two confounders causing biases. Inspired by our causal view and to align LLMs to generate less biased outputs, we leverage the reward model in RL alignment as an instrumental variable to provide interventional feedback. By adaptively weighting biased samples, we propose Causality-Aware Alignment via RL with interventional feedback. Extensive experiments across multiple text generation tasks demonstrate the effectiveness of our method in debiasing LLMs.

Limitations

Our method utilizes the reward model as an instrumental variable to provide interventional feedback. Despite promising results, the reward models used in this paper may not perfectly satisfy the requirements for being an instrumental variable. As discussed in Section 3.2, an ideal reward model to serve as an instrumental variable in our Causality-Aware Alignment should not inherit biases from its training data or input prompts. This requirement aligns with the practical challenge of training an unbiased, high-quality reward model for RL finetuning of LLMs. We leave further studies on the reward model for future work.

Ethical Consideration

Our proposed method aims to account for causality and utilize interventional feedback in RL alignment for LLM debiasing. There are potential risks in introducing additional biases from poorly designed or low-quality reward models. These biases may be reinforced during RL finetuning, leading to the misalignment of LLMs. Thus extra caution is needed when applying this method for practical use.

References

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. Semeval 2017 task 10: Scienceie - extracting keyphrases and relations from scientific publications. *CoRR*, abs/1704.02853.

Ashutosh Baheti, Ximing Lu, Faeze Brahman, Ronan Le Bras, Maarten Sap, and Mark Riedl. 2024. Improving language models with advantage-based offline policy gradients. In *The Twelfth International Conference on Learning Representations*.

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Carlos Brito and Judea Pearl. 2002. Generalized instrumental variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, page 85–93, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec

Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Boxi Cao, Hongyu Lin, Xianpei Han, Fangchao Liu, and Le Sun. 2022. Can prompt probe pretrained language models? understanding the invisible risks from a causal view. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5796–5808, Dublin, Ireland. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. 2019. Jigsaw unintended bias in toxicity classification.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019, Toronto, Canada. Association for Computational Linguistics.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe RLHF: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun SHUM, and Tong Zhang. 2023. RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Anoop Kadan., Manjary P. Gangan, Deepak P., and Lajish V. L. 2022. Towards an enhanced understanding of bias in pre-trained neural language models: A survey with special emphasis on affective bias. In *Responsible Data Science*, pages 13–45, Singapore. Springer Nature Singapore.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. 2023. Pretraining language models with human preferences. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17506–17533. PMLR.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Shuai Li, Zhao Song, Yu Xia, Tong Yu, and Tianyi Zhou. 2023. The closeness of in-context learning and weight shifting for softmax regression. *arXiv preprint arXiv:2304.13276*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xu Liu, Tong Yu, Kaige Xie, Junda Wu, and Shuai Li. 2024. Interact with the explanations: Causal debiased explainable recommendation system. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 472–481.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with

human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Richard Yuanzhe Pang and He He. 2021. Text generation by learning from demonstrations. In *International Conference on Learning Representations*.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Baolin Peng, Linfeng Song, Ye Tian, Lifeng Jin, Haitao Mi, and Dong Yu. 2023. Stabilizing rlhf through advantage model and selective rehearsal. *arXiv preprint arXiv:2309.10202*.

Alexander Peysakhovich and Dean Eckles. 2018. Learning causal effects from many randomized experiments using regularized instrumental variables. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 699–707, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In *The Eleventh International Conference on Learning Representations*.

Shaina Raza, Deepak John Reji, and Chen Ding. 2022. Dbias: detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics*, pages 1–21.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos Garea, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6252–6272, Toronto, Canada. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Maximilian Seitzer, Bernhard Schölkopf, and Georg Martius. 2021. Causal influence detection for improving efficiency in reinforcement learning. *Advances in Neural Information Processing Systems*, 34:22905–22918.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural media bias detection using distant supervision with BABE - bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Daniel Cox, Yiming Yang, and Chuang Gan. 2024. SALMON: Self-alignment with principle-following reward models. In *The Twelfth International Conference on Learning Representations*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings*

of Machine Learning Research, pages 35151–35174. PMLR.

Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023. A causal view of entity bias in (large) language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15173–15184, Singapore. Association for Computational Linguistics.

Siyin Wang, Jie Zhou, Changzhi Sun, Junjie Ye, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Causal intervention improves implicit sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6966–6977, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yotam Wolf, Noam Wies, Yoav Levine, and Amnon Shashua. 2023. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*.

Junda Wu, Rui Wang, Tong Yu, Ruiyi Zhang, Handong Zhao, Shuai Li, Ricardo Henao, and Ani Nenkova. 2022. Context-aware information-theoretic causal de-biasing for interactive sequence labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3436–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Junda Wu, Tong Yu, and Shuai Li. 2021. Deconfounded and explainable interactive vision-language retrieval of complex scenes. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2103–2111.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36.

Yu Xia, Fang Kong, Tong Yu, Liya Guo, Ryan A Rossi, Sungchul Kim, and Shuai Li. 2024. Which llm to play? convergence-aware online model selection with time-increasing bandits. *arXiv preprint arXiv:2403.07213*.

Yu Xia, Junda Wu, Tong Yu, Sungchul Kim, Ryan A. Rossi, and Shuai Li. 2023. User-regulation deconfounded conversational recommender system with bandit feedback. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 2694–2704, New York, NY, USA. Association for Computing Machinery.

Canwen Xu, Zexue He, Zhankui He, and Julian McAuley. 2022. Leashing the inner demons: Self-detoxification for language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11530–11537.

Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. 2024. RLCD: Reinforcement learning from contrastive distillation for LM alignment. In *The Twelfth International Conference on Learning Representations*.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. RRHF: Rank responses to align language models with human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. 2020. Interventional few-shot learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 2734–2746. Curran Associates, Inc.

Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, et al. 2023. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint arXiv:2304.05524*.

Ruiyi Zhang and Tong Yu. 2023. Understanding demonstration-based learning from a causal perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1465–1475, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenkai Zhang, Hongyu Lin, Xianpei Han, and Le Sun. 2021. De-biasing distantly supervised named entity recognition via causal intervention. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4803–4813, Online. Association for Computational Linguistics.

Rui Zheng, Shihan Dou, Songyang Gao, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Limao Xiong, Lu Chen, et al. 2023. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241, Toronto, Canada. Association for Computational Linguistics.