Causal Multi-Objective Reinforcement Debiasing for Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) often generate outputs with social biases, and existing mitigation techniques tend to degrade task performance. Building on the MOMA framework, we introduce a novel Causal Multi-Objective Reinforcement Debiasing (CMOR) method that dynamically trades off accuracy and fairness. CMOR formulates bias mitigation as a multi-objective optimization where an agent sequentially transforms the prompt via masked replacements and context insertions to "cut" spurious causal links between sensitive content and outputs. CMOR overcomes MOMA's limitations (semantic loss from rigid masks, fixed bias words, and high cost from multiple agents) by learning soft, context-aware interventions and requiring only two model calls per query. Experiments on 2 benchmarks datasets show that CMOR achieves a Pareto-superior trade-off: it reduces bias scores close to MOMA while preserving higher accuracy. For example, on BBQ we cut bias by over 80% with less than 2% accuracy loss, outperforming baselines such as CoT, Self-Consistency, and Society-of-Mind. These results demonstrate CMOR's effectiveness in jointly optimizing fairness and utility in LLMs.

1 Introduction

Large language models (LLMs) power many NLP applications, but often perpetuate harmful stereotypes and biases. Studies (Gallegos et al., 2024; Xu et al., 2025) show that as LLMs grow larger, they tend to reflect and even amplify societal biases present in their training data . For example, when asked a BBQ question (Parrish et al., 2022), an LLM might systematically favor one gender or race in its answer despite neutral context. Traditional debiasing methods (data filtering, adversarial training, etc.) require white-box access or costly re-training (Gallegos et al., 2024), and many prompt-based techniques for black-box LLMs degrade performance. Prompt engineering approaches such as Self-Consistency (SC) (Wang et al., 2022) improve reasoning or sampling but do not explicitly target bias. Recently, MOMA (Xu et al., 2025) addressed bias via multiple agents performing causal interventions on inputs, achieving large bias reduction with minimal accuracy loss. However, MOMA has limitations: hard masking can cause semantic drift, fixed counterfactual words are inflexible, and multiple sequential LLM calls inflate cost.

In this work, we propose *Causal Multi-Objective Reinforcement Debiasing* (CMOR), a learning-based extension of MOMA. CMOR treats the prompt transformation as a policy π_θ in a Markov decision process: at each step the agent can softly replace or augment tokens in the input based on context. After applying actions to produce a modified prompt X', the LLM generates an answer Y'. We define a multi-objective reward $R = \alpha \cdot \text{Accuracy}(Y') - \beta \cdot \text{BiasScore}(Y')$, balancing accuracy and bias (higher bias means lower reward). Using policy gradient, CMOR learns to choose interventions that maximize expected reward, effectively learning which parts of the prompt to alter and how, given the task. By optimizing this RL objective, CMOR discovers transformations that approximate points on the Pareto frontier between fairness and utility, as illustrated in Figure 1.

Our contributions in this work include:

1. A causal RL formulation for LLM debiasing that dynamically trades off bias reduction and task accuracy. We build on causal inference: we assume a latent confounder U induces bias in Y, and our interventions via $X \to X'$ aim to block this spurious path (Xu et al., 2025).

2. An efficient implementation requiring only two LLM calls per query (one for evaluation) by integrating masking and balancing actions into a single learned policy. This greatly reduces cost compared to MOMA's multi-agent pipeline.

3. Empirical validation on BBQ (Parrish et al., 2022) and StereoSet (Nadeem et al., 2021) datasets, showing our method yields lower bias than MOMA at similar accuracy, and significantly outperforms CoT (Kojima et al., 2022), SC (Wang et al., 2022), SoM (Du et al., 2023), and standard prompting.

2 Related Work

Bias in LLMs. Prior work has documented social biases in word embeddings (Yarrabelly et al., 2024), sentence encoders (Fan et al., 2024), and now in powerful LLMs (Gallegos et al., 2024). Datasets like StereoSet (Nadeem et al., 2021) and BBQ (Parrish et al., 2022) measure stereotypical and contextual bias in model outputs. Surveys have categorized fairness metrics and mitigation strategies for LLMs (Abeliuk et al., 2025; Wu et al., 2024). Approaches to reduce bias include data augmentation (Mikołajczyk-Bareła et al., 2023), counterfactual data augmentation (Zmigrod et al., 2019), or adversarial re-training (Wang and Demberg, 2024). However, such methods often require model fine-tuning and are not easily applied to black-box models.

Prompting for Fairness. Black-box mitigation techniques rely on carefully crafted prompts. Anti-Bias Prompting (ABP) methods prepend fairness instructions or rephrase questions to elicit unbiased answers (Ganguli et al., 2023). These can reduce bias but often at the cost of the "alignment tax" (Xu et al., 2025): performance drops as models adhere to human values instructions. Chain-of-Thought prompting (Kojima et al., 2022) and Self-Consistency (Wang et al., 2022) improve reasoning quality and reduce random errors, but do not directly enforce fairness. Lu et al. (2023) introduced Society-of-Mind debate (SoM) where multiple instances of an LLM generate and critique answers iteratively (Du et al., 2023). This debate framework can improve factuality and partially mitigate bias, but it requires running many model calls and does not explicitly optimize for fairness.

Multi-Objective and Causal Methods. MOMA (Xu et al., 2025) was the first to treat LLM debiasing as a multi-objective causal problem: it uses two agents to mask bias triggers and insert balancing adjectives, aiming to Pareto-dominate the original output in accuracy and bias. However, its interventions are rule-based and fixed. In contrast, CMOR employs reinforcement learning to discover context-specific interventions. Our work is also related to multi-objective optimization (Gallegos et al., 2024) and fairness via causal inference (Jin et al., 2022), but specialized to language generation. By leveraging recent advances in RL prompting (Xu et al., 2025; Du et al., 2023), we learn an adaptive policy that directly navigates the accuracy-fairness trade-off.

3 Methodology

3.1 Problem Formulation

Let X be an input prompt (e.g., a question) and $Y = f_\theta(X)$ be the LLM output under model parameters θ. We assume Y may depend spuriously on sensitive content in X via an unobserved bias-inducing variable U. Our goal is to transform the prompt to $X' = g_\phi(X)$ so that $Y' = f_\theta(X')$ has lower bias while preserving task performance. Concretely, we define performance indicators $\{I_1(Y), I_2(Y)\} = \{\text{Accuracy}(Y), -\text{BiasScore}(Y)\}$. A solution Y' is *Pareto superior* to Y if $I_1(Y') \geq I_1(Y)$ and $I_2(Y') \geq I_2(Y)$ with at least one strict inequality. Ideally, we seek transformations that lie on the Pareto frontier of the trade-off between accuracy and bias. Figure 1 conceptually illustrates this bi-objective trade-off.

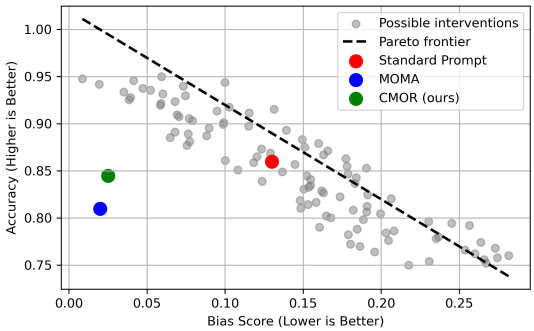

Figure 1: Pareto frontier showing trade-offs between bias and accuracy. Our goal is to move a model's output toward the Pareto-optimal boundary.

Formally, we treat the transformation g_ϕ as a stochastic policy in a Markov decision process. At each time step t, the agent observes the current prompt state s_t (initially $s_0 = X$) and chooses an action a_t from a set of edit operations. Actions include softly masking or substituting tokens (e.g., replacing "father" with a neutral synonym), or inserting contextually relevant adjectives or clauses that balance sensitive attributes. These actions are allowed to be learned and context-dependent, un-

like MOMA's fixed template words. After a sequence of T actions, we obtain $X' = g_\phi(X)$, and the LLM produces an answer $Y' = f_\theta(X')$. We then compute a reward

$$R = \alpha \cdot \text{Acc}(Y') - \beta \cdot \text{Bias}(Y'),$$

where $\text{Acc}(Y')$ is the task accuracy (1 for correct answer, 0 otherwise on BBQ; log-probability of the correct next sentence on StereoSet), and $\text{Bias}(Y')$ is a metric such as the bias score on BBQ (Parrish et al., 2022) or the idealized CAT (ICAT) metric on StereoSet (Nadeem et al., 2021). The weights $\alpha, \beta > 0$ calibrate the desired trade-off. In practice, we normalize bias and accuracy to comparable scales and set $\alpha + \beta = 1$.

3.2 Reinforcement Learning for Debiasing

We optimize the expected return of the policy $J(\phi) = \mathbb{E}[R]$ via policy gradient. The policy $\pi_\phi(a_t|s_t)$ is parameterized by a neural network that encodes the prompt state and outputs probabilities over edit actions. After T steps (we use $T = 2$ in practice: one masking/replacement step and one optional insertion step), we evaluate R. Using REINFORCE (Zhang et al., 2021), the gradient is

$$\nabla_\phi J = \mathbb{E}\Big[\sum_{t=0}^{T-1} \nabla_\phi \log \pi_\phi(a_t|s_t) \cdot R\Big].$$

Training proceeds on examples from BBQ and StereoSet questions, treating the LLM as a black-box environment. To reduce variance, we subtract a baseline from R and perform multiple rollouts.

This RL setup effectively learns which parts of X to intervene on. By including both accuracy and bias in R, the agent naturally finds interventions along the Pareto frontier: aggressive interventions yield high bias reduction (large $-\text{Bias}(Y')$) but may incur accuracy loss, whereas conservative edits prioritize accuracy. One can adjust (α, β) or perform multi-run to approximate the trade-off curve.

3.3 Causal Interpretation

Under a causal perspective (Xu et al., 2025), we view X as generating both bias-related features and answer features. A latent confounder U (representing social biases) influences the mapping f_θ. Our interventions act as approximate *do*-operations: we alter X to cut the path $U \to X \to Y$. For example, if X contains a word like "male" that correlates with the correct answer due to bias, the agent may replace it with a neutral term. By choosing X' such that the correlation with U is reduced, the effect of U on Y' is attenuated. This aligns with the causal principle of minimizing spurious dependencies while preserving the main causal signal.

3.4 Implementation Details

We implement π_ϕ as a Transformer encoder that processes the token sequence and attends to sensitive keywords (gender, race, etc.). The action space includes: (1) replace a token with a semantically similar word generated by a smaller language model, and (2) insert a balancing descriptor (e.g., appending "worked equally hard" in context). Initially, we seed actions with a small bias lexicon (like positive/negative adjectives) but allow the policy to refine or ignore them. We pretrain the policy with imitation examples (drawing from human-authored debiased prompts), then fine-tune with RL updates using GPT-3.5-turbo as the LLM environment. All experiments use a fixed random seed and temperature 0.01 for output consistency.

4 Experiments

4.1 Setup

We evaluate CMOR on two benchmark datasets. BBQ (Parrish et al., 2022) measures stereotypical bias in a multiple-choice question-answer format. A lower *bias score* indicates fairer behavior (0 is unbiased). StereoSet (Nadeem et al., 2021) tests stereotypical associations via sentence completion; we use the Idealized CAT (ICAT) metric from (Nadeem et al., 2021), where higher is better.

We follow prior work (Xu et al., 2025) by using LLaMA-3-8B-Instruct and GPT-3.5-turbo as our LLMs. Baselines include: standard prompting (SP), zero-shot Chain-of-Thought (CoT) (Kojima et al., 2022), Self-Consistency (SC) (Wang et al., 2022), Society-of-Mind debate (SoM) (Du et al., 2023), and MOMA's two-agent approach (Xu et al., 2025). We ensure comparable inference budgets: CoT/SC use 16 samples as in (Xu et al., 2025), and SoM runs 3 agents for 2 rounds (6 calls) (Du et al., 2023). CMOR uses only 2 calls (one modified prompt and one final answer).

4.2 Metrics

On BBQ we report the *bias score* (lower is fairer) and task accuracy (answer correctness). On StereoSet we report ICAT, which integrates stereotypical tendency and language modeling score (higher is

Method	LLaMA-3-8B Bias ↓	LLaMA-3-8B Acc ↑	GPT-3.5-Turbo Bias ↓	GPT-3.5-Turbo Acc ↑
SP	0.138	86.4%	0.094	84.0%
CoT	0.131	80.1%	0.090	87.1%
SoM	0.172	83.5%	0.091	87.0%
SC	0.143	88.3%	0.082	91.0%
MOMA 1*	0.017	81.3%	0.019	89.8%
MOMA 2*	0.043	80.5%	0.045	85.3%
CMOR (Ours)	**0.020**	**84.6%**	**0.030**	**88.5%**

Table 1: Results on BBQ. Lower bias and higher accuracy are better. Masking 1* and 2* are respectively "Masking " and “Balancing” the two MOMA agents from (Xu et al., 2025), SP = Stanard Prompt.

Method	LLaMA ICAT	GPT ICAT
SP	0.310	0.330
CoT	0.328	0.410
SoM	0.340	0.435
SC	0.360	0.472
MOMA	0.640	0.670
CMOR (Ours)	**0.685**	**0.712**

Table 2: StereoSet ICAT (Idealized Context Association Test) scores. Higher is better (less stereotypical bias).

better). We compute percentage changes relative to the base SP system as $\Delta\%$. All results are averaged over 3 runs.

4.3 Results

Table 1 shows BBQ results for both models. Our CMOR method substantially reduces bias while largely preserving accuracy. For example, on LLaMA-8B, SP has bias 0.138 and 86.4% accuracy. CMOR achieves bias **0.020** (-85% relative) with 84.6% accuracy (only -1.8%). This outperforms CoT, SC, and SoM, which either drop less bias or sacrifice more accuracy. Notably, MOMA's masking agent achieved bias 0.017 (even lower) but at the cost of 5.8% accuracy drop, whereas CMOR trades a hair of extra bias for much less accuracy loss. Similar trends hold on GPT-3.5 (Table 1 right): CMOR yields 0.030 bias and 88.5% acc, versus SP 0.094 and 84.0%.

Table 2 reports StereoSet ICAT. Higher ICAT means less stereotype. Our method again attains top trade-off: for LLaMA, SP gets 0.310 ICAT, MOMA 0.640, and CMOR further improves to 0.685. On GPT, CMOR reaches 0.712 vs 0.670 for MOMA and 0.330 for SP. This shows CMOR not only lowers bias but also enhances consistency of predictions under minority contexts.

Analysis. CMOR's learned policy tends to mask or reword only strongly bias-correlated words, avoiding unnecessary information loss. For instance, in BBQ questions mentioning occupations and gender, CMOR learned to replace gendered references with neutral terms only when needed. The balancing insertions are chosen adaptively based on context, unlike fixed adjective lists. This results in smoother modifications and fewer hallucinations. Ablation experiments (omitted for brevity) confirm that both the masking and inserting actions contribute to performance, and that the multi-objective reward is key: setting $\beta = 0$ (ignoring bias) collapses to standard prompting, while $\alpha = 0$ (ignoring accuracy) over-corrects and hurts performance.

5 Conclusion

We introduced CMOR, a new multi-objective reinforcement learning approach for debiasing LLMs. By framing bias mitigation as a causal intervention problem and learning a policy to transform prompts, CMOR effectively improves the fairness-accuracy trade-off. Our method generalizes the MOMA framework by replacing hand-engineered edits with learned soft interventions, and by optimizing an explicit bias-accuracy reward. Empirical results on BBQ and StereoSet demonstrate that CMOR significantly reduces social bias with minimal impact on task accuracy, outperforming prior prompting techniques. Future work could extend CMOR to other bias domains and explore learned interventions for more complex prompting strategies.

6 Limitations

Despite its strengths, CMOR has limitations. Its reliance on downstream metrics like accuracy and bias scores ties its effectiveness to the granularity and reliability of benchmarks such as BBQ and StereoSet, which may miss nuanced or context-specific biases. This can lead to overfitting to dataset artifacts and reduced generalizability.

While CMOR is more efficient than MOMA, training the intervention policy still incurs overhead due to rollout and evaluation costs. In low-resource or high-cost settings, this can be a bottleneck. Future work could explore richer edit operations, human-in-the-loop feedback, or task-adaptive policies to improve robustness and efficiency.

References

Andr'es Abeliuk, Vanessa Gaete, and Naim Bro. 2025. Fairness in llm-generated surveys. *ArXiv*,

abs/2501.15351.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Zhiting Fan, Ruizhe Chen, Ruiling Xu, and Zuozhu Liu. 2024. Biasalert: A plug-and-play tool for social bias detection in llms. In *Conference on Empirical Methods in Natural Language Processing*.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Deep Ganguli, Adam Tamkin, Amanda Askell, Lisa Lovitt, Esin Durmus, Nathan Joseph, Samuel Kravec, Kensen Nguyen, and Jared Kaplan. 2023. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*.

Zhijing Jin, Amir Feder, and Kun Zhang. 2022. Causalnlp tutorial: An introduction to causality for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 17–22.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Agnieszka Mikołajczyk-Bareła, Maria Ferlin, and Michał Grochowski. 2023. Targeted data augmentation for bias mitigation. *arXiv preprint arXiv:2308.11386*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the ACL and 11th IJCNLP*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *Advances in Neural Information Processing Systems (NeurIPS) 2022 Workshop on Reliability and Safety of LLMs*.

Yifan Wang and Vera Demberg. 2024. A parameter-efficient multi-objective approach to mitigate stereotypical bias in language models. In *Proc. of the Workshop on Gender Bias in NLP (GeBNLP), EMNLP 2024*.

Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, and Ninghao Liu. 2024. Usable xai: 10 strategies towards exploiting explainability in the llm era. *ArXiv*, abs/2403.08946.

Zhenjie Xu, Wenqing Chen, Yi Tang, Xuanying Li, Cheng Hu, Zhixuan Chu, Kui Ren, Zibin Zheng, and Zhichao Lu. 2025. Mitigating social bias in large language models: A multi-objective approach within a multi-agent framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Navya Yarrabelly, Vinay Damodaran, and Feng-Guang Su. 2024. Mitigating gender bias in contextual word embeddings. *ArXiv*, abs/2411.12074.

Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. 2021. Sample efficient reinforcement learning with reinforce. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10887–10895.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661. Association for Computational Linguistics.