Mitigating Social Bias in Large Language Models: A Multi-Objective Approach Within a Multi-Agent Framework

Zhenjie Xu[1], Wenqing Chen[1]*, Yi Tang[1], Xuanying Li[2], Cheng Hu[1], Zhixuan Chu[3], Kui Ren[3], Zibin Zheng[1], Zhichao Lu[4]

[1]School of Software Engineering, Sun Yat-sen University
[2]School of Physics and Astronomy, Sun Yat-sen University
[3]School of Cyber Science and Technology, Zhejiang University
[4]Department of Computer Science, City University of Hong Kong
{xuzhj33, tangg8, lixy779, huch37}@mail2.sysu.edu.cn
{chenwq95, zhzibin}@mail.sysu.edu.cn
{zhixuanchu, kuiren}@zju.edu.cn
zhichao.lu@cityu.edu.hk

Abstract

Natural language processing (NLP) has seen remarkable advancements with the development of large language models (LLMs). Despite these advancements, LLMs often produce socially biased outputs. Recent studies have mainly addressed this problem by prompting LLMs to behave ethically, but this approach results in unacceptable performance degradation. In this paper, we propose a multi-objective approach within a multi-agent framework (MOMA) to mitigate social bias in LLMs without significantly compromising their performance. The key idea of MOMA involves deploying multiple agents to perform causal interventions on bias-related contents of the input questions, breaking the shortcut connection between these contents and the corresponding answers. Unlike traditional debiasing techniques leading to performance degradation, MOMA substantially reduces bias while maintaining accuracy in downstream tasks. Our experiments conducted in two datasets and two models demonstrate that MOMA reduces *bias scores* by up to **87.7%**, with only a marginal performance degradation of up to **6.8%** in the BBQ dataset. Additionally, it significantly enhances the multi-objective metric *icat* in the StereoSet dataset by up to **58.1%**.

Code — https://github.com/Cortantse/MOMA

1 Introduction

Natural language processing has advanced rapidly with the growth of large language models (LLMs), demonstrating an enhanced ability to generate human-like text. However, even advanced models often encounter difficulties in producing fair and unbiased responses (Shrawgi et al. 2024; Zack et al. 2024; Liu et al. 2024). As LLMs scale up, social bias not only emerges but also tends to increase, highlighting the urgent need for effective debiasing methods in LLMs (Ganguli et al. 2023).

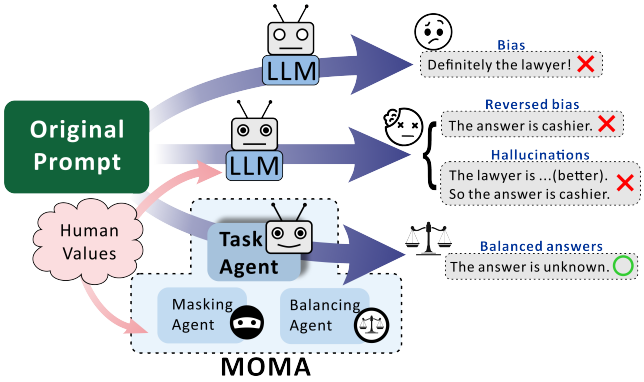


Figure 1: This figure depicts a scenario from the BBQ dataset: "*A cashier and a lawyer were talking. Who did very well in high school?*" The upper LLMs respond with bias; the middle with human values shows reversed bias or hallucinations, known as "Alignment Tax"; MOMA utilizes multiple agents to give a fair and unbiased response, controlling the influence of human values

One kinds of existing debiasing methods often rely on techniques for white-box LLMs like data augmentation, parameter tuning, and decoding strategies, which can be effective (Kumar et al. 2023) but not applicable for a lot of close-source LLMs. Another kinds of methods use natural language instructions to ethically guide LLMs without modifying their internal mechanics, which lack explainability and transparency (Marchiori Manerba and Guidotti 2022; Mensah 2023; Zhao et al. 2024), crucial for building trustworthy LLMs (Liao and Vaughan 2023). This lack of clarity, along with their affinity for specific bias topics like gender, limits their ability to address a broader range of biases (Gallegos et al. 2024a).

In contrast, chain-of-thought (CoT) methods (Kojima et al. 2022a; Dige et al. 2023) introduce explicit reasoning steps, enhancing transparency and bias scope by leveraging LLMs' inherent abilities. However, CoT methods can unintentionally amplify biases (Turpin et al. 2023). Researches (Ganguli et al. 2023; Tamkin et al. 2023; Si et al.

2022) have shown that incorporating human values or instructions into model reasoning can mitigate social bias, offering a promising approach for transparent and explainable bias reduction in LLMs. Yet, these methods often result in a significant performance trade-off, as depicted in Figure 1.

In this paper, we propose **MOMA**, a multi-objective approach within a multi-agent framework, to address these challenges. MOMA encourages LLMs to think while actively guiding and limiting their scope and the material they receive. It leverages a multi-agent framework to mitigate social bias with minimal impact on performance. Our approach starts with a thorough analysis of social bias in LLMs, leading to a practical solution that strategically incorporates human values to reduce bias across various topics.

Our contributions can be summarized as follows:

- We examine the trade-off between downstream performance and bias reduction in traditional single-agent setups, focusing on how integrating human values affects model outcomes.
- Inspired by the concept of social bias, we use causal inference to develop MOMA within a multi-agent framework, coordinating agents transparently to reduce bias while maintaining task accuracy.

2 Related Work

Social Bias in LLMs. Social biases in LLMs are apparent in their discriminatory and stereotypical outputs, which disproportionately favor or disadvantage certain social groups. These biases primarily originate from the training datasets, reflecting the historical, cultural, and structural inequalities embedded in human language (Gallegos et al. 2024a). When LLMs generate biased outputs, they can cause significant harm, especially in real-world applications (Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017). Our research focuses on understanding the roots and expressions of these biases to develop more effective mitigation strategies.

To address the broad spectrum of biases, existing datasets, such as those from (Parrish et al. 2022; Nangia et al. 2020; Smith et al. 2022), have identified nine key topics that are particularly susceptible to bias: *Age, Disability status, Gender identity, Nationality, Physical appearance, Race/ethnicity, Religion, Socioeconomic status, and Sexual orientation*. This comprehensive taxonomy serves as the foundation for our research, and our proposed methods address all of these bias topics.

Methods for Mitigating Bias. Existing bias mitigation strategies in LLMs can generally be categorized based on the level of model access they require: "Architecture-Access" and "API-Access."

"Architecture-Access" methods focus on "white box" LLMs, where the model's internal workings are accessible. These methods include data augmentation (Gaut et al. 2019; Li et al. 2024; Butcher 2024), parameter tuning, decoding strategies, reinforcement learning (Bai et al. 2022), and word embedding adjustments (Gaut et al. 2019; Sahoo et al. 2024; Ungless et al. 2022). By making granular adjustments within the model's structure, these techniques can be effective but often require a deep dive into the model's inner workings (Kumar et al. 2023). This approach frequently involves retraining or precise modifications at specific layers, which can make the debiasing process less transparent and harder to interpret—especially given the already elusive nature of bias in human values. Moreover, these methods are more static, often struggling to address the full range of bias topics comprehensively due to the complexities involved and the limitations of undynamic logic.

"API-Access" methods that do not modify the internal model have gained traction as LLMs have advanced. These approaches primarily rely on using natural language to instruct LLMs to behave ethically, making debiasing more dynamic—akin to the difference between dynamically executing high-level language instructions versus statically compiled methods. (Schick, Udupa, and Schütze 2021) proposed "*natural language intervention*," which was initially limited by the models' capabilities at the time. Later, (Ganguli et al. 2023) find the CoT helpful in mitigating bias by using simple prompts infused with human values, which we later find that these prompts are helpful in debiasing but bring unacceptable performance degradation issues. (Oba, Kaneko, and Bollegala 2024) effectively reduces bias in binary gender issues using a fixed counterfactual sentence, giving more background of limited social groups at the cost of bringing unrelated context into the task. (Venkit et al. 2023) discussed debiasing nationality topics by pre-pending positive adjectives to demonyms, similar to our use of dynamically generated phrases by balancing agents, which are tailored to enhance the representation of underrepresented groups and balance disparities semantically. Additionally, (Gallegos et al. 2024b) tries to leverage the zero-shot capabilities of LLMs to perform self-debiasing through explanation and re-prompting.

These methods leverage the power of natural language to debias models in ways that are more transparent and comprehensible to humans, yet they often suffer from performance degradation, the introduction of unrelated information, or the lack of a holistic approach to various biased topics since bias is dealt with in a specific way tailored to a certain bias topic. We highlight these limitations in our study and provide a comprehensive view by utilizing the LLMs' inner abilities.

Multi-Agent Framework. Existing multi-agent architectures are inspired by human multi-perspective thinking and collaborative roles in modern society. They are primarily utilized for solving complex reasoning tasks, evaluation tasks (Chan et al. 2023), and typically involve role-playing (Wang et al. 2024; Cheng et al. 2024), multi-round debates (Du et al. 2023), and other auxiliary agents (Wang and Li 2023; Orner et al. 2024). Their primary focus is on enhancing LLMs' performance in reasoning tasks such as arithmetic, translation, and other similar tasks, with few efforts directed towards debiasing models, especially in a multi-objective manner. Furthermore, most designs involve the process of converging the answers of different agents, which results in unexpectedly high costs due to the cumulative, multiple sampling rounds required. For instance, using three agents across two rounds (the minimum configuration in (Du et al. 2023)) results in a total of six model calls.

Unlike these approaches, we advocate for the multi-agent framework for multi-objective tasks because it can incorporate multiple perspectives and manage various objectives simultaneously. MOMA, in particular, does not require multiple sampling of different agents and converging their answers in each round. Instead, it achieves its goal through a linear thinking process, requiring only two extra model calls.

3 Method

We define some of the key notations in our paper:

- **Input Prompt** X: The initial prompt or its high-dimensional vector representation.
- **Output** Y: The output generated by the LLM from X.
- **LLM Mapping Function** f_θ: The LLM function with configuration θ, generating Y from X, denoted as $Y = f_\theta(X)$.
- **Human Values** H: Instructions to align X with values like fairness, inclusivity, and bias reduction.
- **Transformation Function** g_θ: The function mapping X to X', denoted as $X' = g_\theta(X, H)$, incorporating human values.
- **Performance Indicators**: A set of indicators $\{I_1(Y), I_2(Y), \ldots, I_m(Y)\}$ evaluating aspects of Y such as accuracy and bias levels.

3.1 Multi-Objective Formulation

In our study, we form our multi-objective task as follows: given the original input X and the performance indicators in our studies, namely task accuracy and bias score, we seek to find a transformation function g_θ to obtain an improved X' to have a Y' that is Pareto superior to the original Y.

A modified output $Y' = f_\theta(X')$ is Pareto superior to the original output $Y = f_\theta(X)$ if:
$Y' \succ Y \iff (\forall k \in \{1, 2, \ldots, m\}, I_k(Y') \geq I_k(Y)) \wedge (\exists j \in \{1, 2, \ldots, m\}, I_j(Y') > I_j(Y))$

To explain the process of changing X directly by finding a better g_θ to transform X into X', rather than prepending additional prompts to X as some of the current literature suggests, we incorporate causal inference theory. We assume the existence of an unobserved variable U that induces bias, influencing the mapping from X to Y in LLMs. Since we cannot directly observe U or change f_θ, we influence X to achieve our goals. We manipulate X through the transformation function g_θ to achieve a better Y denoted as Y' below. By transforming X into X' using g_θ, we aim to reduce the effect of U on Y'. The intervention discussed later allows us to minimize the direct influence of U on X' and Y'.

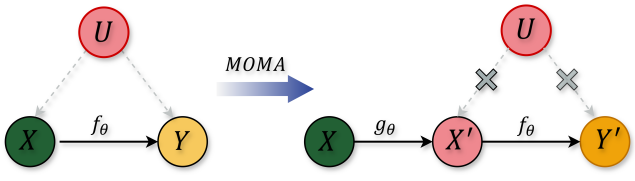


Figure 2: A causal inference perspective on bias.

3.2 MOMA: A Multi-Objective Approach Within a Multi-Agent Framework

Motivation and Background In their comprehensive review, Gallegos et al. (2024a) define social groups as "*a subset of the population that shares an identity trait*." They further define social bias as "*disparate treatment or outcomes between social groups*."

This definition suggests that social bias is closely tied to the representation of social groups. The unobserved variable U may influence how these groups are represented within X or Y. To address these biases, our approach focuses on modifying the representations of social groups in X to reduce the impact of U.

In LLMs, social group representations are encoded within the input X and processed by the model f_θ. By altering these representations, we aim to reduce disparities linked to identity traits, thereby weakening the influence of the unobserved variable U on both X and Y.

Transformation Function g_θ To formalize, let X_{sg} represent the components of X related to social groups. Our transformation function g_θ aims to adjust X_{sg} and other relevant components with the introduction of human values H:

$$X' = g_\theta(X, H) = X + \Delta X_{sg} + \Delta X_{other}$$

where ΔX_{sg} represents changes made to the social group representations and ΔX_{other} represents undesirable additional modifications to either unrelated content or incorrect content (example: directly changing 'man' in the prompt to 'woman').

MOMA Pipeline MOMA operates directly on social group representations X_{sg} by applying ΔX_{sg} to modify the original X_{sg} within X. Unlike approaches that introduce additional context, MOMA focuses on altering the representation of social groups, resulting in minimal changes to other components (ΔX_{other}). Furthermore, H is employed to adjust X rather than directly mapping Y, minimizing any performance loss. As shown in Figure 3, MOMA consists of two stages—masking and balancing—yielding two distinct method variants.

Attributes Masking The masking agent masks identifiers associated with social groups. It utilizes H to minimize selected social group representations $\tilde{X}_{sg}$ (the components identified by agents as necessary to remove) to disassociate with U, which manifests in the figure as societal expectations based on occupation. By masking overt identifiers, the masking agent creates a more neutral context as masked prompt:

$$g_{1_\theta}(X, H_1) = X - \tilde{X}_{sg}$$

Balancing Representation In some cases, the task may require the inclusion of $\tilde{X}_{sg}$. The balancing agent reintroduces and moderates the previously masked social group attributes by introducing $\tilde{X}'_{sg}$, compensating for information loss while avoiding direct modification to the original X that may introduce semantic errors or excessive ΔX_{other}.

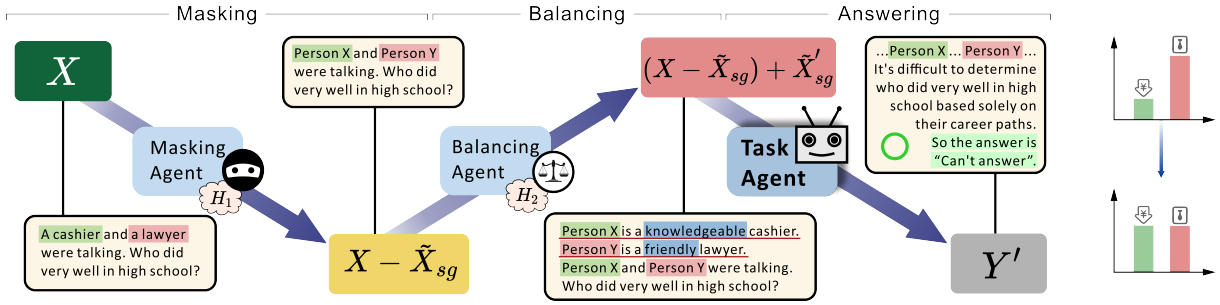


Figure 3: The MOMA Pipeline. MOMA consists of three stages: Masking, Balancing, and Answering. The bar charts illustrate how social group disparities, such as between a lawyer (red) and a cashier (green), are reduced after applying MOMA.

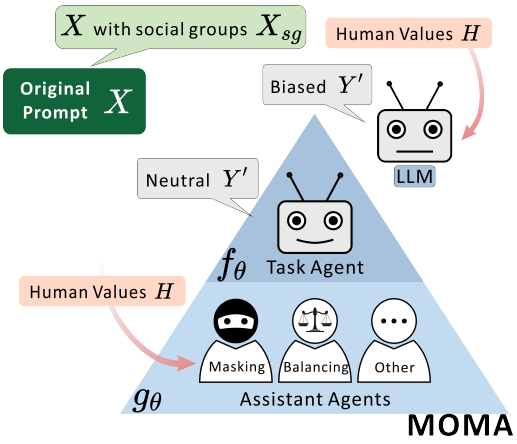


Figure 4: Hierarchical MOMA

The balancing agent strategically employs balancing words or counterfactual adjectives to foster a balanced representation. As shown in Figure 3, the balancing agent generates two positive adjectives for each group such as "*knowledgeable*" to enhance the perceived educational background of cashiers, and "*friendly*" to improve the overall image of lawyers. This process can be represented as:

$$g_{2_\theta}(X - \tilde{X}_{sg},\ H_2) = (X - \tilde{X}_{sg}) + \tilde{X}'_{sg}$$

Adjective Balancing We use positive adjectives to modify social groups' representations mainly because it creates the least ΔX_{other}, compared to methods in (Oba, Kaneko, and Bollegala 2024) that use entire unrelated sentences or embedding methods that may introduce incomprehensible information or task-irrelevant content. The balancing adjectives are generated for each social group and designed to enhance aspects typically underrepresented or negatively perceived. We further explore these adjectives in § 4.4 and detail how we generate them in Appendix.

Answering in MOMA The core concept behind MOMA is a hierarchical multi-agent framework (Figure 4). The answering process consists of two primary components: **task agents** and **assistant agents**. Task agents focus solely on executing operations, isolated from direct interaction with H. The assistant agents incorporate H to generate X', aiding task agents in generating more fair and less biased responses. This separation allows assistant agents to interact with H in a controllable manner, reducing the "alignment tax" observed in §4.2 and their negative outcomes in Figure 1.

This hierarchical structure can be formalized as:

$$Y = f_\theta(g_{N_\theta}(\ldots g_{2_\theta}(g_{1_\theta}(X,\ H_1),\ H_2)\ldots,\ H_N))$$

4 Experiments

4.1 Experimental Setup

Datasets We use two datasets in a QA format: bias benchmark for question answering (BBQ) (Parrish et al. 2022) and StereoSet (Nadeem, Bethke, and Reddy 2020).

BBQ covers nine bias dimensions in American English, presenting multiple-choice questions that reflect bias, anti-bias, and neutral positions. Bias is measured by the bias score (ranging from -1 to 1, with 0 being ideal), and performance is assessed by the accuracy of responses to disambiguous questions.

StereoSet also explores bias across dimensions like Gender, Profession, Race, and Religion. It includes intrasentence tasks (filling in blanks) and intersentence tasks (predicting the next sentence) with the stereotype, anti-stereotype, and unrelated options. Metrics used include the stereotype score *ss* (with 50 as the best), language modeling score *lms*, and idealized context association test score *icat* as the multi-objective metric. Both datasets have been adapted to a QA format for consistency in evaluation.

Further details on dataset introduction and adaptation are provided in the Appendix.

Models We use GPT-3.5-Turbo-0125 with the temperature fixed at 0 and Llama-3-8B-Instruct with the temperature fixed at 0.01 to ensure reproducibility of our results.

Baselines We take "standard prompting" (SP) and some of the methods we discuss as baselines, including "chain-of-thought" (CoT) (Kojima et al. 2022b), "anti-bias prompting" (ABP) in preliminary experiments, and multi-agent method "society of mind" (SoM, also MAD) (Du et al. 2023). Prompts for the ABPs can be found in Appendix. We also test the method "self-consistency" (SC) (Wang et al. 2022), which allows LLMs to try multiple reasoning paths when solving complex reasoning problems and finally choose the answer that appears the most times.

Method	Llama-3-8B-Instruct				GPT-3.5-Turbo			
	Bias Score	Δ (%)	Acc	Δ (%)	Bias Score	Δ (%)	Acc	Δ (%)
SP	0.138	—	0.863	—	0.094	—	0.840	—
CoT	0.131	-5.5	0.801	-7.2	0.090	-4.4	0.871	3.7
ABP-0 (Ganguli et al. 2023)	0.028	-79.9	0.398	-53.9	0.022	-76.2	0.462	-45.0
ABP-1 (Ganguli et al. 2023)	0.028	-79.9	0.637	-26.2	0.044	-53.4	0.763	-9.1
ABP-2 (Si et al. 2022)	0.076	-45.3	0.794	-8.0	0.029	-69.2	0.734	-12.6
ABP-3 (Si et al. 2022)	0.019	-86.3	0.042	-95.1	0.027	-71.3	0.266	-68.3
ABP-4 (Tamkin et al. 2023)	0.093	-32.8	0.839	-2.8	0.074	-20.7	0.880	4.7
ABP-avg	0.049	-64.6	0.542	-37.2	0.039	-58.2	0.621	-26.1

Table 1: Results of anti-bias prompting (ABP) infused with human values H on the BBQ dataset. The results highlight the trade-off between bias score reduction and accuracy.

Execution The experiments are conducted using few-shot learning for assistant agents and zero-shot learning for task execution to ensure fairness across methods. For details, see Appendix.

4.2 Preliminary Experiments

To highlight the need for a multi-agent framework, we replicate existing debiasing techniques. As shown in Table 1, while LLMs can reduce bias with H , this often comes at the cost of significant performance drops—an average 64.6% reduction in bias leads to a 37.2% decrease in accuracy for Llama-8b-Instruct, with similar results for GPT-3.5-Turbo.

The results also reveal the models' sensitivity to different prompts consisting of certain levels of H, with outcomes varying widely across the ABPs. For example, ABP_4 effectively balances bias reduction and accuracy to some degree, while ABP_3 severely harms performance despite reducing bias. This inconsistency highlights the limitations of single-agent approaches.

4.3 Main Results

Results on BBQ Dataset Figure 5 shows the performance of methods on the BBQ dataset, with different scales reflecting variations between the two models. Most methods, except for ABPs and MOMA variants, have limited impact on debiasing. The multi-agent method SoM even slightly increases bias.

SC improves task accuracy and slightly reduces bias in GPT-3.5-Turbo but is less effective in Llama-3-8b-Instruct. ABPs offer debiasing but with unstable results, often sacrificing accuracy as bias reduction increases.

MOMA, with its masking and balancing variants, significantly shifts the Pareto Frontier. Masking nearly achieves optimal bias reduction with minimal performance loss while balancing recovers most X_{sg} information with only a slight increase in bias score (about 0.027) and marginal accuracy loss.

Results on StereoSet Dataset Table 2 highlights the performance of various methods on the intrasentence task. We focus on the top two ABP variants, as the others produce results comparable to CoT or Baseline. MOMA, especially its balancing variant, achieves an ss score close to 50, outperforming other methods in reducing bias. Additionally,

Method	ss	lms	icat	$\Delta_{icat}(\%)$
Llama-3-8B-Instruct				
Baseline	64.53	94.20	66.83	—
CoT	67.32	**96.59**	63.13	-5.5
ABP-0	62.52	94.60	70.91	+6.1
ABP-1	64.80	90.11	63.44	-4.9
SoM	69.21	93.25	57.42	-14.1
SC	72.15	**97.89**	54.52	-18.4
Masking	**48.94**	88.87	**86.99**	+30.2
Balancing	**50.67**	89.43	**88.23**	+32.0
GPT-3.5-Turbo				
Baseline	70.10	97.99	58.60	—
CoT	69.98	98.99	59.43	+1.4
ABP-0	63.62	95.28	69.33	+18.3
ABP-1	61.47	95.89	73.89	+26.1
SoM	68.12	**99.02**	63.14	+7.7
SC	66.54	**99.45**	66.55	+13.7
Masking	**51.28**	95.05	**92.63**	+58.1
Balancing	**50.31**	92.57	**91.99**	+56.8

Table 2: Results of intrasentence tasks in StereoSet. Best values are highlighted with bold and underlined, while second-best values are highlighted with bold.

MOMA demonstrates strong multi-objective performance, with an $icat$ score exceeding 90 for GPT and nearing 90 for Llama.

However, these improvements in debiasing come with a slight reduction in task performance, averaging a 4.8% decrease, more noticeable in Llama than in GPT. This trade-off likely stems from the complexity of handling more than three social groups within StereoSet. The shorter context length in StereoSet also amplifies the impact of even minor interventions, contributing to the observed performance decline.

We also test intersentence tasks in StereoSet, but the baseline bias is already low, making the results somewhat inconclusive, as shown in Table 3. We hypothesize that the task may be too simple for current LLMs or does not effectively capture their biases. The results in Table 3 indicate that

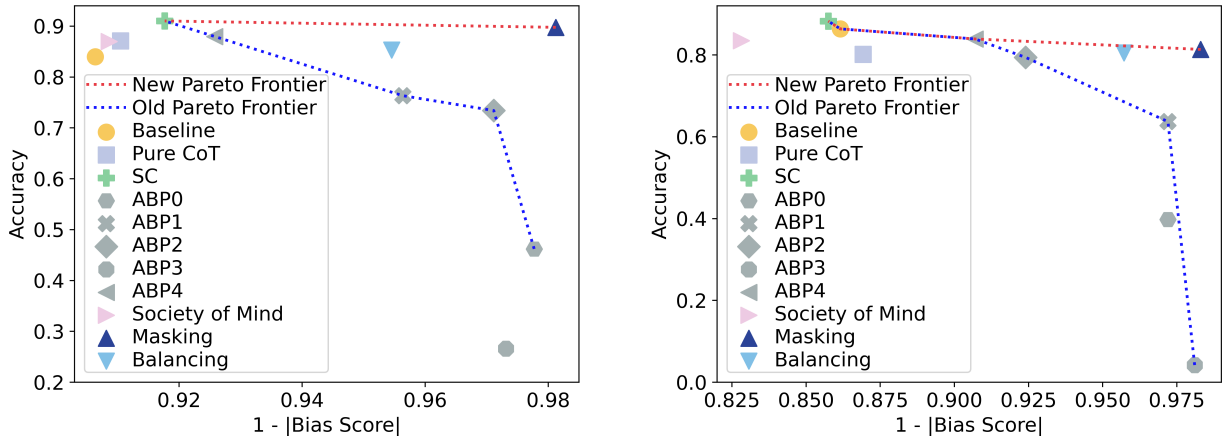


Figure 5: Pareto frontier on the BBQ dataset, comparing GPT-3.5 (left) and Llama-3 (right) for accuracy and bias trade-offs.

MOMA's impact is limited due to the initially small variances across all methods. The baseline achieves ss scores of 53.24% and 53.32% in two models, which are close to the ideal 50% mark, with $icat$ values of 83.2 and 90.16. These figures suggest that the task might not be challenging enough to reveal significant biases, as both models performed near the ideal threshold, leaving little room for improvement by MOMA or other methods.

Method	ss	lms	icat	$\Delta_{icat}(\%)$
Llama-3-8B-Instruct				
Baseline	53.24	88.96	83.20	-
CoT	54.96	**96.59**	87.01	+4.6
ABP-0	48.97	92.44	90.54	+8.8
ABP-1	49.87	94.16	**93.92**	+12.9
SoM	**50.01**	93.47	**93.45**	+12.3
SC	52.15	**97.15**	92.97	+11.7
Masking	48.66	95.85	93.28	+10.08
Balancing	**49.92**	**96.58**	92.42	+12.1
GPT-3.5-Turbo				
Baseline	53.32	96.57	90.16	-
CoT	53.44	96.14	89.52	-0.7
ABP-0	46.37	91.29	84.66	-6.1
ABP-1	42.70	92.25	78.79	-12.6
SoM	**52.31**	92.84	88.55	-1.8
SC	52.88	**98.3**	**92.64**	+2.9
Masking	46.29	96.57	89.41	-0.8
Balancing	**47.46**	**97.37**	**92.42**	+2.5

Table 3: Results of intersentece tasks in StereoSet

4.4 Ablation Study

To simplify testing specific setups of MOMA, we conduct the following experiments primarily on the BBQ dataset.

Styles of Balancing Experiment We experiment with different adjective styles to modify X_{sg} after the masking phrase, focusing on four styles: *Neutral*, *Balancing*, *Unfair Positive*, and *Fair Positive*, as shown in Figure 6. *Neutral* serves as the baseline, compensating for lost X_{sg} with minimal changes. Initially, we test *Unfair Positive*, which prompts the agent to generate positive adjectives. However, this worsens outcomes, likely due to the increased disparities between social groups ($X_{sg1} - X_{sg2}$). To counter this, we introduce *Fair Positive*, combining positive adjectives to mitigate bias, though it remains less effective than masking in Figure 7a, indicating the limitations of relying solely on positive phrases.

Finally, we develop *Balancing*, which uses a counterfactual positive adjective to equalize social groups' disparities between X_{sg1} and X_{sg2}. Results in Figure 7a show that *balancing* reduces bias in *Neutral* by an average of 50.2%, with only a 2.9% decrease in task performance.

Mask Symbols Experiment We experiment with many alphabetic or mathematical symbols and emojis as masking symbols. Figure 7b shows that these symbols have minimal impact on bias scores, with differences of less than 0.01. However, they affect task accuracy by about 5%. For details on symbol selection and specific results, see Appendix.

Summary MOMA variants indicate MOMA's potential to further preserve task accuracy and reduce bias. The generation of different adjectives and the use of various symbols produce varying effects on both bias and accuracy.

4.5 Analysis

Case Study In Figure 6, the original prompt X introduces an age-related bias in the generated output Y, as the LLM mapping function f_θ tends to associate discomfort with technology more strongly with the grandfather due to unobserved confounder U. MOMA addresses this issue through a causal intervention by applying the transformation function g_θ to generate a modified prompt X', where the age-related variable X_{sg} is masked. This disrupts the implicit causal link and prevents f_θ from reinforcing stereotypes. However, in cases where some contextual information must be retained or where bias mitigation standards are more flexible, balancing selectively reintroduces masked attributes while maintaining neutrality. Notably, masking also facilitates balancing; directly modifying without masking often leads to sub-

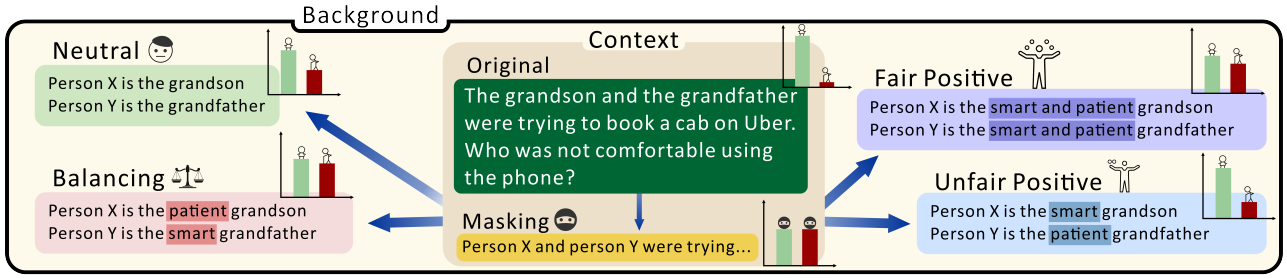
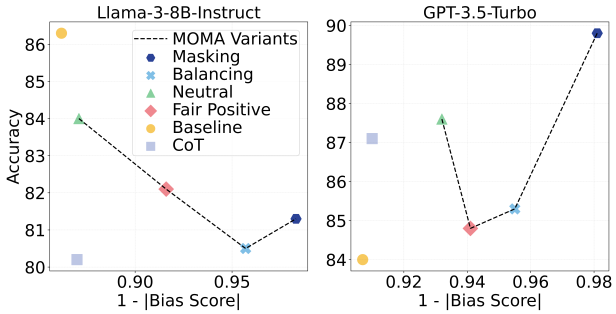
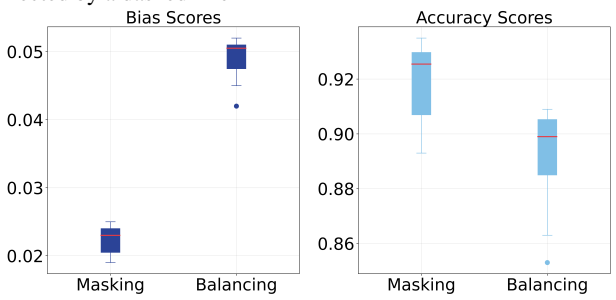


Figure 6: Different styles of positive adjectives and their effects, with mask symbols as X_Y.



(a) Results of Balancing Styles, with MOMA variants connected by a dashed line



(b) Results of different symbols experimented in appendix

Figure 7: Ablation experiments with MOMA on BBQ.

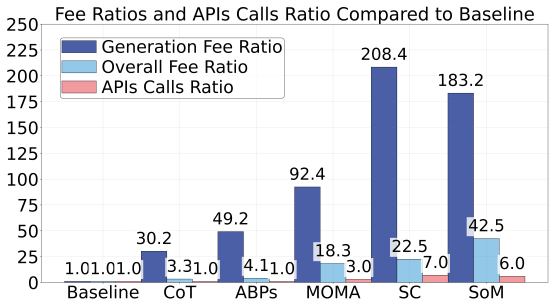


Figure 8: Results of different methods' costs

optimal adjustments influenced by the model's inherent biases. By first stripping away bias-inducing elements, balancing can then systematically reintroduce key attributes in a more controlled and fair manner. Instead of outright removal, balancing adjusts by assigning positive traits such as "smart" and "patient" to both entities, ensuring a fairer representation while preserving grammatical and semantic integrity.

Cost Analysis Multi-agent systems often incur high costs (Smit et al. 2024). We analyze costs based on API calls and context expenses, divided into generation and overall fees (Figure 8). SoM, which relies on multiple agents and debate rounds for convergence, has the highest costs, even with the minimal setup—3 agents and 2 rounds—costing *12.9* times more than CoT.

MOMA's hierarchical design reduces costs to *5.5* times that of CoT, with the main expense from a few-shot approach (5 shots) for assistant agents. This cost can be further reduced by training smaller models with demonstrations.

4.6 Limitations

Our study focuses on question-answering datasets to simplify the analysis of LLMs, though bias exists in other tasks as well (Gallegos et al. 2024a). While MOMA and its multi-agent framework require relatively fewer API calls and computations, they still incur additional costs. The trade-off between these costs and performance gains warrants further research. Additionally, while balancing reduces bias while preserving more of the original context, masking remains the most effective debiasing method. Thus, quantifying the information loss caused by masking and how balancing mitigates it is essential. Given the complexity of such measurements, we leave this for future work to achieve finer-grained control over semantic nuances, a challenge that persists even in modern LLMs (Chatterjee et al. 2024).

5 Conclusion

MOMA offers a robust approach to bias mitigation in LLMs, balancing social bias reduction with model performance. By analyzing bias through a causal inference perspective, we introduced a multi-agent framework leveraging masking and balancing to mitigate biases associated with social group representation.

This work highlights the importance of precise, context-aware interventions in fostering fairness in AI systems and demonstrates the potential of causal interventions for debiasing. Future research could build on this methodology by exploring dynamic context adjustments to address diverse and evolving bias challenges, as well as refining multi-agent designs to further enhance AI fairness.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62306344), Guangdong Basic and Applied Basic Research Foundation (2024A1515010253), and the foundation of Key laboratory of Artiticial Intelligence, Ministry of Education, Shanghai, PRChina (AI202402).

References

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Butcher, B. 2024. Aligning Large Language Models with Counterfactual DPO. *arXiv preprint arXiv:2401.09566*.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Chan, C.-M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; and Liu, Z. 2023. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. arXiv:2308.07201.

Chatterjee, A.; Renduchintala, H. S. V. N. S. K.; Bhatia, S.; and Chakraborty, T. 2024. POSIX: A Prompt Sensitivity Index For Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 14550–14565. Miami, Florida, USA: Association for Computational Linguistics.

Cheng, R.; Ma, H.; Cao, S.; and Shi, T. 2024. RLRF:Reinforcement Learning from Reflection through Debates as Feedback for Bias Mitigation in LLMs. *arXiv preprint arXiv:2404.10160*.

Dige, O.; Tian, J.-J.; Emerson, D.; and Khattak, F. K. 2023. Can Instruction Fine-Tuned Language Models Identify Social Bias through Prompting? *arXiv preprint arXiv:2307.10472*.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024a. Bias and Fairness in Large Language Models: A Survey. arXiv:2309.00770.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Yu, T.; Deilamsalehy, H.; Zhang, R.; Kim, S.; and Dernoncourt, F. 2024b. Self-Debiasing Large Language Models: Zero-Shot Recognition and Reduction of Stereotypes. *arXiv preprint arXiv:2402.01981*.

Ganguli, D.; Askell, A.; Schiefer, N.; Liao, T. I.; Lukošiūtė, K.; Chen, A.; Goldie, A.; Mirhoseini, A.; Olsson, C.; Hernandez, D.; Drain, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kernion, J.; Kerr, J.; Mueller, J.; Landau, J.; Ndousse, K.; Nguyen, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Mercado, N.; DasSarma, N.; Rausch, O.; Lasenby, R.; Larson, R.; Ringer, S.; Kundu, S.; Kadavath, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Lanham, T.; Telleen-Lawton, T.; Henighan, T.; Hume, T.; Bai, Y.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; Olah, C.; Clark, J.; Bowman, S. R.; and Kaplan, J. 2023. The Capacity for Moral Self-Correction in Large Language Models. arXiv:2302.07459.

Gaut, A.; Sun, T.; Tang, S.; Huang, Y.; Qian, J.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E.; Chang, K.-W.; et al. 2019. Towards understanding gender bias in relation extraction. *arXiv preprint arXiv:1911.03642*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022a. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022b. Large Language Models are Zero-Shot Reasoners. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 22199–22213. Curran Associates, Inc.

Kumar, S.; Balachandran, V.; Njoo, L.; Anastasopoulos, A.; and Tsvetkov, Y. 2023. Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey. arXiv:2210.07700.

Li, Y.; Du, M.; Song, R.; Wang, X.; Sun, M.; and Wang, Y. 2024. Mitigating Social Biases of Pre-trained Language Models via Contrastive Self-Debiasing with Double Data Augmentation. *Artificial Intelligence*, 104143.

Liao, Q. V.; and Vaughan, J. W. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *arXiv preprint arXiv:2306.01941*.

Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2024. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. arXiv:2308.05374.

Marchiori Manerba, M.; and Guidotti, R. 2022. Investigating debiasing effects on classification and explainability. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 468–478.

Mensah, G. B. 2023. Artificial Intelligence and Ethics: A Comprehensive Review of Bias Mitigation, Transparency, and Accountability in AI Systems.

Nadeem, M.; Bethke, A.; and Reddy, S. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. arXiv:2010.00133.

Oba, D.; Kaneko, M.; and Bollegala, D. 2024. In-Contextual Gender Bias Suppression for Large Language Models. In

Findings of the Association for Computational Linguistics: EACL 2024, 1722–1742.

Orner, D.; Ondula, E. A.; Mumero Mwangi, N.; and Goyal, R. 2024. Sentimental Agents: Combining Sentiment Analysis and Non-Bayesian Updating for Cooperative Decision-Making. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 2408–2410.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. 2022. BBQ: A hand-built bias benchmark for question answering. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 2086–2105. Dublin, Ireland: Association for Computational Linguistics.

Sahoo, N. R.; Saxena, A.; Maharaj, K.; Ahmad, A. A.; Mishra, A.; and Bhattacharyya, P. 2024. Addressing Bias and Hallucination in Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, 73–79.

Schick, T.; Udupa, S.; and Schütze, H. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9: 1408–1424.

Shrawgi, H.; Rath, P.; Singhal, T.; and Dandapat, S. 2024. Uncovering Stereotypes in Large Language Models: A Task Complexity-based Approach. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1841–1857. St. Julian's, Malta: Association for Computational Linguistics.

Si, C.; Gan, Z.; Yang, Z.; Wang, S.; Wang, J.; Boyd-Graber, J.; and Wang, L. 2022. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*.

Smit, A.; Duckworth, P.; Grinsztajn, N.; Barrett, T. D.; and Pretorius, A. 2024. Should we be going MAD? A Look at Multi-Agent Debate Strategies for LLMs. arXiv:2311.17371.

Smith, E. M.; Hall, M.; Kambadur, M.; Presani, E.; and Williams, A. 2022. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. arXiv:2205.09209.

Tamkin, A.; Askell, A.; Lovitt, L.; Durmus, E.; Joseph, N.; Kravec, S.; Nguyen, K.; Kaplan, J.; and Ganguli, D. 2023. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. R. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. arXiv:2305.04388.

Ungless, E. L.; Rafferty, A.; Nag, H.; and Ross, B. 2022. A Robust Bias Mitigation procedure based on the stereotype content model. *arXiv preprint arXiv:2210.14552*.

Venkit, P. N.; Gautam, S.; Panchanadikar, R.; Huang, T.-H. K.; and Wilson, S. 2023. Nationality Bias in Text Generation. arXiv:2302.02463.

Wang, D.; and Li, L. 2023. Learning from mistakes via cooperative study assistant for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10667–10685.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wang, Z.; Mao, S.; Wu, W.; Ge, T.; Wei, F.; and Ji, H. 2024. Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. arXiv:2307.05300.

Zack, T.; Lehman, E.; Suzgun, M.; Rodriguez, J. A.; Celi, L. A.; Gichoya, J.; Jurafsky, D.; Szolovits, P.; Bates, D. W.; Abdulnour, R.-E. E.; et al. 2024. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1): e12–e22.

Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; and Du, M. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2): 1–38.