Mitigating Bias in LLMs via EquiSync: A Multi-Objective Optimization Perspective

Anonymous ACL submission

Abstract

Natural Language Processing (NLP) has seen remarkable advancements in Large Language Models (LLMs). Despite these advancements, a persistent challenge remains: LLMs often produce biased outputs. This paper introduces EquiSync, a novel method designed to mitigate social bias in LLMs without significantly compromising their performance. EquiSync utilizes a multi-agent framework, incorporating three agents that employ a two-phase approach: Attributes Masking and Attributes Balancing. This method aligns with human values transparently and reduces disparities between social groups. Unlike traditional debiasing techniques, which often lead to performance degradation, EquiSync achieves substantial bias reduction while maintaining or even improving accuracy in downstream tasks. Our experiments demonstrate that EquiSync 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 **56.98%**. The code will be made publicly available upon acceptance.

1 Introduction

Natural language processing has advanced rapidly with the growth of large language models (LLMs) like GPT-4, demonstrating an enhanced ability to generate human-like text (OpenAI et al., 2024). However, even these 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).

Existing debiasing methods often rely on implicit adjustments such as data augmentation, parameter tuning, decoding strategies, etc. These techniques can be effective (Kumar et al., 2023). But they sometimes fall short in terms of explainability and transparency (Marchiori Manerba and Guidotti, 2022; Mensah, 2023; Zhao et al., 2024), which are the crucial elements in building trustworthy LLMs (Liao and Wortman Vaughan, 2024). In contrast, Chain-of-Thought (CoT) methods (Kojima et al., 2022a; Dige et al., 2023) introduce explicit reasoning steps, enhancing transparency but can unintentionally amplify biases (Turpin et al., 2023b). In response, research such as (Ganguli et al., 2023; Tamkin et al., 2023; Si et al., 2022) has demonstrated that incorporating human values or instructions and then engaging models in thinking can effectively mitigate social bias, offering a promising avenue for transparent and explainable bias mitigation in LLMs. However, we later observe that these methods often significantly decrease performance, presenting a critical trade-off issue, see Figure 1.

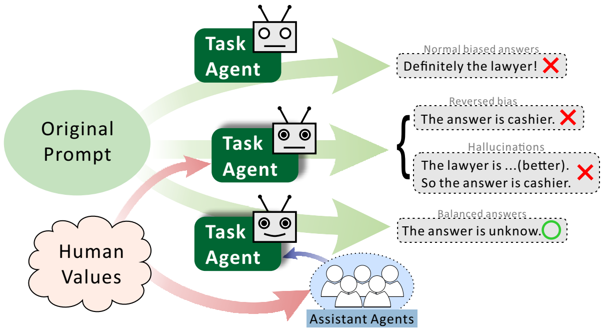

Figure 1: This figure depicts a scenario: "*A cashier and a lawyer were talking. Who did very well in high school?*" The first agent responds with bias; the second with human values shows reversed bias or hallucinations; our methods utilize multiple agents to give a fair and unbiased response, controlling the influence of human values.

In this paper, we propose a novel method called **EquiSync**, which addresses these challenges through a multi-agent framework designed to mitigate social bias without compromising performance. Our approach begins with a comprehensive analysis of social bias and its origins in LLMs, moving towards a practical solution that incorporates human values strategically to reduce bias. Our work contributes the following:

- We examine the trade-off between downstream performance and bias reduction in traditional single-agent setups, focusing on how the integration of human values influences model outcomes.
- To optimize the use of human values and control their influence more effectively, we develop a multi-agent framework to achieve multi-objective optimization.
- Inspired by the definition of social bias, we introduce the EquiSync method within our proposed framework. EquiSync orchestrates multiple agents to collaborate effectively, each with specialized roles and focused objectives.

2 Related Work

Social Bias in LLMs. Social bias in LLMs manifests through discriminatory patterns and stereotypical representations that unfairly favor or disadvantage certain social groups. This bias primarily stems from the training datasets, which inherently reflect the historical, cultural, and structural inequalities in human language use (Gallegos et al., 2024a). Consequently, biased outputs from LLMs can lead to significant harm when these models are employed in real-world contexts (Bolukbasi et al., 2016; Caliskan et al., 2017). Addressing these biases is crucial, especially given the widespread application of LLMs.

Recognizing the diverse manifestations of bias, datasets such as those developed by (Parrish et al., 2022; Nangia et al., 2020; Smith et al., 2022) categorically highlight nine main attributes prone to bias: *Age, Disability status, Gender identity, Nationality, Physical appearance, Race/ethnicity, Religion, Socioeconomic status, and Sexual orientation*. These datasets play a vital role in quantifying and understanding biases in models, providing a comprehensive taxonomy that guides our research to address and encompass all identified facets of bias systematically.

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."

The former focuses on the "white box" LLMs; methods include data augmentation (Gaut et al., 2019; Li et al., 2024b; Butcher, 2024), parameter tuning, decoding strategies, reinforcement learning (Bai et al., 2022), word embedding adjustment (Gaut et al., 2019; Sahoo et al., 2024; Ungless et al., 2022), etc. Adjusting at a granular level within the model's structure, these techniques are sometimes effective but require a deep dive into the model's inner workings (Kumar et al., 2023), often involving retraining or precise adjustments at specific layers. This makes the debiasing process less transparent and complicates its understanding.

While direct model manipulation methods are prevalent, complementary strategies that do not alter the internal model have also gained traction. (Schick et al., 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 useful in debiasing but have brought unacceptable performance degradation. (Oba et al., 2024) effectively reduced bias in binary gender issues using a fixed counterfactual sentence. (Venkit et al., 2023) discussed debiasing nationality topics by prepending positive adjectives to demonyms, similar to our use of dynamically generated phrases by balancing agents tailored to enhance the representation of underrepresented groups and balance disparities. Additionally, (Gallegos et al., 2024b) leverages 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 a lack of holistic approach to various biased topics. We highlight these limitations in our study and provide a comprehensive view.

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, 2023a; Orner et al., 2024).

For instance, in the research conducted by (Wang and Li, 2023a), a Study Assistant agent is designed to interact with the main LLM to help it learn from incorrect cases. With its simple two-agent design, their system has improved the performance of the main LLM on the BBQ dataset by collecting data and retrieving cases. While their system shares similarities with ours in its hierarchical structure, their primary focus is improving downstream performance, not debiasing models.

Unlike these approaches, we advocate that the multi-agent framework is suited for multi-objective tasks, particularly because it can incorporate multiple perspectives and manage various objectives simultaneously.

3 Method

3.1 Multi-Objective Formulation

The challenge of balancing multiple objectives has long been acknowledged across various systems. Our multi-objective formulation concerns two criteria: **reducing social bias** and **maintaining downstream performance**.

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.*" We adopt these definitions for our study, focusing our methods primarily on balancing these disparities that arise from differences in identity traits. However, we acknowledge that the concept of social bias is dynamic and continually evolving; biases or stereotypes toward certain groups may vary across different times and contexts. Our methods specifically target balancing social groups instead of mitigating certain social biases to ensure flexibility and adaptability.

Reducing Social Bias: Reducing social bias entails strategically aligning model outputs with human values, which serve as proxies for varying definitions of bias. This black-box optimization alignment configures our agents to interpret and adapt to a spectrum of human values. These values act as dynamic standards for bias identification and mitigation. Our proposed framework, detailed in Section 3.2, and its implementation, described in Section 3.3, ensures that agents remain flexible and responsive to diverse and evolving social norms, representing the underlying values defining biases.

Maintaining downstream Performance: Ensuring the maintenance of downstream performance is critical for the functionality and utility of models, as the debiasing process must balance bias mitigation with performance preservation to maintain the model's practical applicability in real-world scenarios.

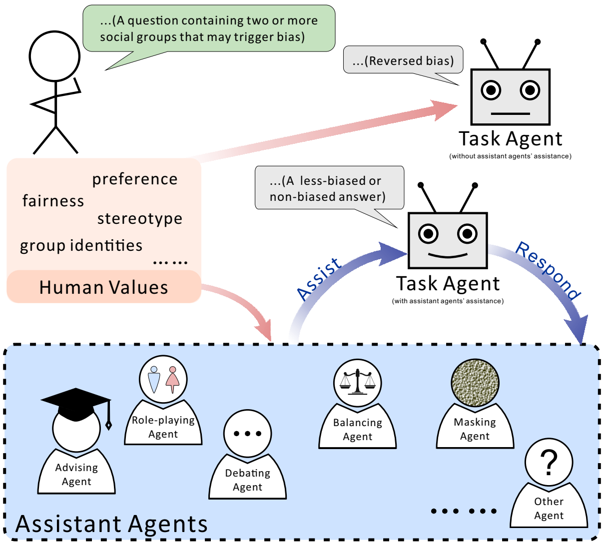

Figure 2: The proposed multi-agent framework.

3.2 Multi-Agent Framework for Debiasing

Human beings and LLMs both encounter large amounts of data infused with biases and inequalities. However, unlike LLM, humans can achieve relatively unbiased responses through contemplation, integrating multiple perspectives, and employing various cognitive regions to process complex information simultaneously or sequentially.

The challenge of producing unbiased responses is pronounced in LLMs. Although LLMs can simulate human-like contemplation through explicit reasoning techniques such as CoT (Wei et al., 2022) and other even more sophisticated methods (Yao et al., 2024) (Wang et al., 2022), they still exhibit biases in their reasoning processes (Turpin et al., 2023a). The contrast between autoregressive models using a left-to-right generation approach and the dynamic non-linear thought processes typical of humans is particularly striking. The linear generation process of such models can lead to compounding biases, as each subsequent word is selected based on the narrow context set by the preceding word, often ignoring broader or conflicting perspectives that may counteract the bias. This is particularly true for autoregressive models with a left-to-right generation approach. The sequential nature of these models contrasts with the dynamic and nonlinear thought processes typical of humans. This linear generation process may lead to compounding biases because each subsequent word is chosen based on a narrow context set by the preceding words, often overlooking broader or conflicting perspectives that might counteract biases. Consequently, without proper intervention, LLMs can reinforce existing biases as they elaborate on their reasoning, creating a feedback loop that exacerbates these biases.

To counteract this problem, we propose a multi-agent framework to mirror human cognitive abilities in assimilating and processing multiple viewpoints. This framework integrates several candidate agents drawn from existing studies and our own research, each functioning as a debiasing module to intervene in the reasoning process. By utilizing natural language for coordination, these agents collectively work to ensure that the generation of responses aligns with and actively promotes human values concerning bias mitigation.

The multi-agent framework we propose consists of two essential parts: **task agents** and **assistant agents**. While the Task agents are solely responsible for executing operations, intentionally isolated from direct engagement with human values, the assistant agents incorporate human values to aid the task agents in generating fairer and less biased responses. At first glance, this division might appear unnecessary. However, as detailed in Section 4.2, we observe that LLMs, while capable of debiasing themselves when instructed to do so, sometimes suffer from unacceptable performance degradation. This phenomenon, known as the "Alignment Tax," refers to the costs in performance or unintended negative outcomes, such as reverse bias or increased erroneous outputs like hallucinations, when models are overly aligned with specific human values (see Figure 1). By dividing tasks between agents and assistant agents, our framework regulates these influences and enhances response equity.

Figure 2 presents a multi-agent framework that leverages the collective intelligence of varied agents. The roles of some assistant agents are conceptualized as follows:

- **Perspective Expansion:** Agents provide a range of viewpoints, engage in advising, debating (Du et al., 2023; Wang et al., 2024), memorizing and recalling (Wang and Li, 2023b), and role-playing (Li et al., 2024a; Cheng et al., 2024) to enrich understanding.
- **Contextual Focus:** Agents help refocus the attention of task agents away from content that may evoke bias, steering them towards the context. This is achieved through strategic masking of attributes associated with social groups. More details are provided in § 3.3.
- **Social Group Balancing:** Agents reduce disparities among social groups and balance these disparities by modifying their representations in the context. More details are provided in § 3.3.

To our knowledge, this paper is the first work that introduces a multi-agent system framework applied specifically to debiasing. While multi-agent systems have been previously explored, their use in bias mitigation is novel. This research extends the multi-agent framework to effectively address both bias mitigation and performance preservation, enhancing the fairness and utility of AI systems.

3.3 EquiSync

EquiSync implements our proposed multi-agent framework to strategically mitigate social bias by managing variations in social group attributes. This system operationalizes the concept of social bias as differential treatment based on these attributes, and introduces a progressive two-phase approach: Attributes Masking followed by Attributes Balancing.

In the Attributes Masking phase, the Masking Agent masks identifiers associated with social groups. This neutralizes potentially biased prompts, such as societal expectations based on occupation in Figure 3. By masking overt identifiers, the Masking Agent forces the Task Agent to evaluate the scenario based on a neutral context, avoiding reliance on stereotypical perceptions of social groups and thus achieving Contextual Focus.

Following the masking process, the Balancing Agent reintroduces and moderates the previously masked social group attributes to compensate for information loss. It strategically employs balancing words or adjectives before mentioning these groups to foster a balanced representation. For example, as shown in Figure 3 , the Balancing Agent generates adjectives such as "*knowledgeable*" to enhance the perceived educational background of cashiers, and "*friendly*" to improve the overall image of lawyers, effectively using LLMs' stereotypes to counter stereotypes—cultivating non-toxic fruits from unhealthy soil.

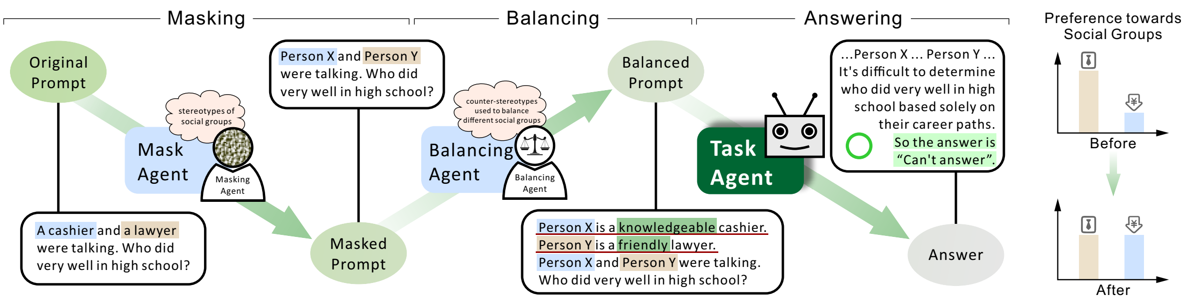

Figure 3: The EquiSync Pipeline —— The Masking Agent begins by masking social attributes in the prompt to reduce bias, causing some information loss. The Balancing Agent then compensates for this loss, ensuring equitable representation of social groups. The Assistant agents promote human values by adjusting the Task Agent's perception of different social groups. This coordination prevents the Task Agent's direct interaction with human values, thus achieving multi-objective.

The balancing adjectives are varied and tailored to each social group, designed to enhance aspects typically underrepresented or negatively perceived. The agent first compares the attributes of the targeted social group with those of opposite groups. It then generates one or multiple adjectives for the targeted group, focusing on improving underrepresented traits rather than merely amplifying positive aspects, as amplifying only positive traits can sometimes inadvertently reinforce stereotypes. We explore these nuances and their implications in § 4.4.

4 Experiments

4.1 Experimental Setup

Datasets We assess the debiasing capabilities of LLMs using two specialized datasets in a question-answering format: BBQ (Parrish et al., 2022) and StereoSet (Nadeem et al., 2020).

BBQ comprises multiple-choice questions across nine social bias dimensions, reflecting bias, anti-bias, and neutral positions within the American English context. We measure bias using the absolute value of the Bias Score, which varies between -1 and 1. Performance is evaluated by the accuracy on disambiguous questions to separate bias detection from logical reasoning capabilities.

StereoSet explores stereotypes in Gender, Profession, Race, and Religion through questions formatted with stereotype, anti-stereotype, and unrelated options. For our purposes, we transform these into a QA format and apply metrics like Stereotype Score*(ss)*, Language Modeling Score*(lms)*, and Idealized CAT Score*(icat)* for comprehensive bias and performance analysis.

Further details on dataset adaptation and metric application 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 set to 0.01 to ensure reproducibility of our results.

Baselines We take Standard Prompting (SP) and some of the methods we discuss as baselines, including CoT (Kojima et al., 2022b), Anti-bias Prompting (ABP) in preliminary experiments, and Multi-agent method Society of Mind (SoM) (Du et al., 2023). Prompts for the ABP methods can be found in Appendix A.x.

Execution We implement the EquiSync method as follows: The assistant agents are few-shot to execute tasks related to masking and balancing. For the masking task, agents are instructed to filter potentially biased content while preserving operational flexibility. In the balancing task, agents are prompted to respect each group and use positive adjectives, integrating human values into their responses. For general task execution, whether by individual or multiple agents, we use zero-shot to ensure fairness in different methods.

4.2 Preliminary Experiments

To substantiate the need for a multi-agent framework, we replicated existing natural language debiasing techniques for contemporary LLMs. The results presented in Table 1 indicate that while LLMs can effectively align with ethical prompts and reflect human values, this alignment often results in considerable reductions in general downstream task performance. Specifically, we observed a 64.6% reduction in the average bias score and a decrease in the average performance score to 37.2% in models such as Llama-8b-Instruct and similar results in GPT-3.5-Turbo. Furthermore, our findings indicate that existing LLMs are extremely sensi-

Method	Llama-3-8B-Instruct Bias Score	$\Delta_b(\%)$	Task Acc.	$\Delta(\%)$	GPT-3.5-Turbo Bias Score	$\Delta(\%)$	Task Acc.	$\Delta(\%)$
SP	0.138	-	0.863	-	0.09362	-	0.83993	-
CoT	0.131	-5.5	0.801	-7.2	0.08954	-4.4	0.87116	3.7
Debias methods (using Explicit Debias Prompt)								
ABP-0	0.028	-79.9	0.398	-53.9	0.02228	-76.2	0.46208	-45.0
ABP-1	0.028	-79.9	0.637	-26.2	0.04365	-53.4	0.76334	-9.1
ABP-2	0.076	-45.3	0.794	-8.0	0.02885	-69.2	0.73407	-12.6
ABP-3	0.019	-86.3	0.042	-95.1	0.02687	-71.3	0.26613	-68.3
ABP-4	0.093	-32.8	0.839	-2.8	0.07421	-20.7	0.87968	4.7
ABP-avg	0.049	-64.6	0.542	-37.2	0.03917	-58.2	0.62106	-26.1

Table 1: Comparison of different Anti-Bias Prompts infused with human values on Llama-3-8B-Instruct and GPT-3.5-Turbo, showing the trade-off between reduced bias and task accuracy.

tive to the nuances of human values, with different prompts yielding contrasting results. For instance, some ABPs like ABP_4 —— *"Really don't discriminate"* mitigate bias effectively at certain levels with minimal performance sacrifice; however, their bias scores remain elevated. Thus, these results substantiate that single-agent approaches are insufficient to handle the complexities of human values in LLMs, driving us to develop methods that address this multi-objective issue more effectively.

4.3 Main Results

Results on the BBQ Dataset Figure 4 and 5 illustrate the performance of two models in the BBQ dataset. EquiSync approaches the optimal performance point more closely than other methods. And it significantly mitigates bias in the BBQ dataset without substantially sacrificing accuracy. Furthermore, EquiSync maintains efficient downstream performance; it only slightly reduces task accuracy in the Llama-3-8B-Instruct model. In contrast, some ABPs experience severe degradation in downstream performance. Notably, in the GPT-3.5-Turbo model, EquiSync not only preserves but also slightly enhances downstream task accuracy, unlike other debiasing methods that offer limited debiasing effects or significantly compromise downstream capabilities.

Results on the StereoSet Dataset The performance comparisons of different debiasing methods on two models are detailed in Table 2 and 3. For Intrasentence Tasks, EquiSync exhibits the most robust debiasing performance, achieving nearly a 50 *ss* score in balancing. This is complemented by a more than 90 *icat* score for GPT and close to 90 *icat* score for Llama, indicating strong multi-objective utility. However, these results are accompanied by a slight decrease in overall model performance. This decline may be attributed to the complexity of managing multiple social groups present in the StereoSet, which sometimes exceed three. This complexity challenges the models' ability to maintain performance while effectively balancing across a broader range of identity traits.

For Intersentence Tasks, the effect of EquiSync is not that significant in the GPT-3.5-Turbo model due to the minor initial variances of the results of all methods in this section. Nevertheless, EquiSync still improves icat by at least 2% in such tasks. Unlike the CoT method, which remains similar to the Baseline, or the ABP methods, which show degraded multi-objective debiasing performance. In Llama-3-8B-Instruct, EquiSync also shows strong debiasing ability. Again, due to the small number of parameters in this model, the effect of our method is not as stable as it is in GPT-3.5-Turbo.

Method	ss	lms	icat
Intrasentence Tasks			
Baseline	70.10	97.99	58.60
CoT	69.98	**98.99**	59.43
ABPα	63.62	95.28	69.33
ABPβ	**61.47**	95.89	73.89
SoM	68.12	99.02	63.14
Masking	51.28	95.05	**92.63**
Balancing	50.31	92.57	**91.99**
Intersentence Tasks			
Baseline	53.32	96.57	90.16
CoT	53.44	96.14	89.52
ABPα	46.37	91.29	84.66
ABPβ	42.70	92.25	78.79
SoM	**52.31**	92.84	88.55
Masking	46.29	96.57	89.41
Balancing	47.46	**97.37**	**92.42**

Table 2: GPT-3.5-Turbo on StereoSet

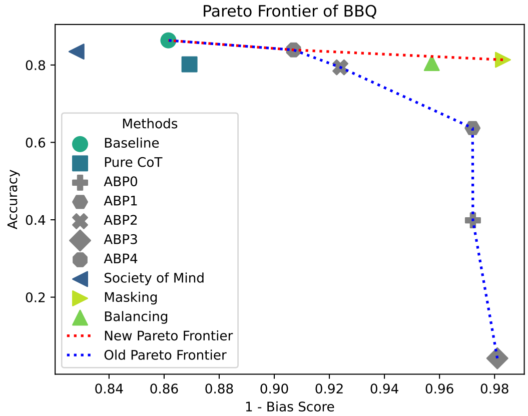

Figure 4: Performance of Llama-3-8b-Instruct on BBQ datasets, our methods effectively flattening the Pareto frontier, indicating robust debiasing and maintained performance.

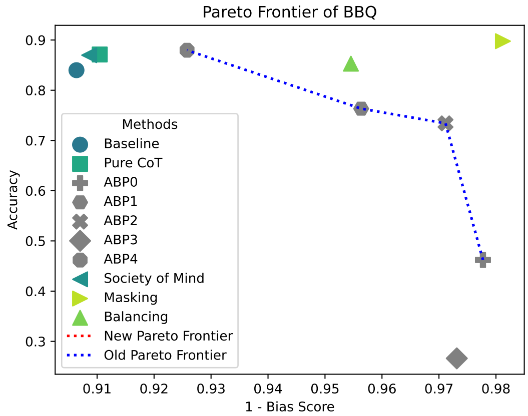

Figure 5: Performance of GPT-3.5-Turbo on BBQ datasets. With Masking, the model achieves optimal performance.

Method	ss	lms	icat
Intrasentence Tasks			
Baseline	64.53	94,20	66.83
CoT	67.32	**<u>96.59</u>**	63.13
ABPα	62.52	94.60	70.91
ABPβ	64.80	90.11	63.44
SoM	69.21	93.25	57.42
Masking	48.94	88.87	86.99
Balancing	50.67	89.43	**<u>88.23</u>**
Intersentence Tasks			
Baseline	53.24	88.96	83.20
CoT	54.96	**<u>96.59</u>**	87.01
ABPα	48.97	92.44	90.54
ABPβ	49.87	94.16	**<u>93.92</u>**
SoM	**<u>50.01</u>**	93.47	93.45
Masking	48.66	95.85	93.28
Balanicng	49.92	96.58	92.42

Table 3: Llama-3-8B-Instruct on StereoSet

Method	Bias Score	Accuracy
Llama-3-8B-Instruct		
Masking	**<u>0.017</u>**	81.3%
Balancing	0.043	80.5%
Neutral	0.129	**<u>84.0%</u>**
Positive	0.084	82.1%
GPT-3.5-Turbo		
Masking	**<u>0.019</u>**	**<u>89.8%</u>**
Balancing	0.045	85.3%
Neutral	0.068	87.6%
Positive	0.059	84.8%

Table 4: Balancing Styles Experiment on BBQ

Symbols	Bias Score	Accuracy	Bias Score	Accuracy
	Masking		Balancing	
X_Y	0.019	0.898	0.045	0.853
Y_X	0.025	0.893	0.042	0.881
α_β	0.020	0.926	0.047	0.897
β_α	0.023	0.935	0.051	0.909
I_II	0.024	0.902	0.050	0.863
II_I	0.024	0.932	0.051	0.899
Cat1	0.022	0.926	0.052	0.906
Cat2	0.023	0.922	0.052	0.899
Smile1	0.020	0.931	0.051	0.906
Smile2	0.025	0.925	0.049	0.903
Average	0.025	0.915	0.050	0.898

Table 5: Mask Symbols Experiment with GPT-3.5-Turbo.

4.4 Ablation Study

In this section, we conducted several key ablation experiments on the proposed EquiSync method.

Styles of Balancing Experiment We conducted experiments to assess the Balancing Agent, using two styles: **neutral** and **positive**. In the neutral style, the agent generates neutral background information for masked social group attributes, while in the positive style, it adds positive prefixes to the masked attributes. Results shown in Table 4 indicate that masking achieves the lowest Bias Score, suggesting that fewer bias-inducing details lead to lower bias. However, masking also results in lower downstream task accuracy compared to methods that include some background information, highlighting the necessity of a Balancing Agent to maintain task performance. The Balancing method, despite slightly reducing downstream task capability due to feature interchange, shows the strongest debiasing effect, underscoring its effectiveness despite some limitations.

Mask Symbols Experiment In the EquiSync method, the Mask Agent needs to mask social group attributes present in the original input using symbols that do not contain any social group information. After fixing the list of symbols, the Mask Agent assigns these symbols to the positions in the input where social group attributes first appear. We conduct an ablation study on the mask symbols. This ablation study focuses on two aspects: the selection of mask symbols and their order of appearance. So we conduct this experiment by selecting additional pairs of mask symbols and further swap their positions in masking processes. The experimental data are shown in Table 5. We can easily found that selecting different symbols and altering their sequence do not significantly impact EquiSync's performance on the BBQ dataset.

4.5 Analysis

This section analyzes why the EquiSync method excels in multi-objective debiasing. As discussed in Section 2, biases in LLMs originate from biases present in the training datasets towards certain social groups. To prevent LLMs from overly focusing on social group attributes and consequently generating biased outputs, we first employ a Mask Agent to obscure features that may lead to bias. Subsequently, a Balancing Agent performs balancing masked group attributes and restores the lost information. The processed text by these auxiliary agents is then fed into the Task Agent. The effectiveness of the EquiSync method in achieving multi-objective debiasing stems from its multi-agent design, which avoids the trade-offs inherent in single-agent systems. The debiasing capability is derived from the Mask operation, which obscures bias-inducing group features, and the Balancing operation, which attaches suitable descriptions to the attributes. Furthermore, the ability to maintain performance in downstream tasks is ensured by the Mask operation, making LLMs focus more on the context and the Balancing operation, restoring background information to prevent a decline in task performance.

4.6 Limitations

Our study selects datasets with question-answering formats to simplify the analysis of LLMs' behavior and effectively measure bias and downstream performance. However, it is important to note that bias manifests across various other tasks as well (Gallegos et al., 2024a). Further, Our methods involve utilizing few-shots for our agents to perform these tasks. We acknowledge that generating high-quality data and training smaller, specialized models could yield more efficient and robust results. We leave this for future work to better equip agents with relevant capabilities and enhance the generalizability of our findings.

5 Conclusion

This study underscores the challenges and trade-offs inherent in debiasing LLMs. While explicit debiasing prompts are instrumental in reducing biases, they can inadvertently impair performance on general downstream tasks due to the added complexity and caution they introduce in model responses. Our proposed EquiSync method leverages a sophisticated multi-agent framework to address these challenges, offering a novel approach to achieving balanced debiasing objectives.

Through extensive experiments on the BBQ and StereoSet datasets, we demonstrated that EquiSync not only effectively mitigates bias but also preserves—sometimes even enhances—the accuracy of downstream tasks. This is achieved without the need for direct model retraining, instead employing a strategy of prompt engineering and dynamic adjustment of model parameters through a multi-agent setup.

EquiSync stands out by creating an environment where models can produce non-toxic outputs from less-than-ideal data conditions. This approach mirrors natural human reasoning processes more closely than traditional methods, promoting fairness and accuracy simultaneously. By integrating these techniques, EquiSync sets a new standard for ethical AI, ensuring LLMs act responsibly in real-world applications without compromising their utility.

As we move forward, it is crucial to continue refining these techniques, exploring their applications in broader contexts, and enhancing their effectiveness across diverse datasets and scenarios. This will help in realizing the full potential of LLMs as tools for positive impact in society.

References

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

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Bradley Butcher. 2024. Aligning large language models with counterfactual dpo. *arXiv preprint arXiv:2401.09566*.

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

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate.

Ruoxi Cheng, Haoxuan Ma, Shuirong Cao, and Tianyu Shi. 2024. Rlrf:reinforcement learning from reflection through debates as feedback for bias mitigation in llms. *arXiv preprint arXiv:2404.10160*.

Omkar Dige, Jacob-Junqi Tian, David Emerson, and Faiza Khan Khattak. 2023. Can instruction fine-tuned language models identify social bias through prompting? *arXiv preprint arXiv:2307.10472*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024a. Bias and fairness in large language models: A survey.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, and Franck Dernoncourt. 2024b. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. *arXiv preprint arXiv:2402.01981*.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Christopher Olah, Jack Clark, Samuel R. Bowman, and Jared Kaplan. 2023. The capacity for moral self-correction in large language models.

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, et al. 2019. Towards understanding gender bias in relation extraction. *arXiv preprint arXiv:1911.03642*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022a. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022b. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Language generation models can cause harm: So what can we do about it? an actionable survey.

Tianlin Li, Xiaoyu Zhang, Chao Du, Tianyu Pang, Qian Liu, Qing Guo, Chao Shen, and Yang Liu. 2024a. Your large language model is secretly a fairness proponent and you should prompt it like one. *arXiv preprint arXiv:2402.12150*.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, Mingchen Sun, and Ying Wang. 2024b. Mitigating social biases of pre-trained language models via contrastive self-debiasing with double data augmentation. *Artificial Intelligence*, page 104143.

Q. Vera Liao and Jennifer Wortman Vaughan. 2024. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*, (Special Issue 5). Https://hdsr.mitpress.mit.edu/pub/aelql9qy.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2024. Trustworthy llms: a survey and guideline for evaluating large language models' alignment.

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

George Benneh Mensah. 2023. Artificial intelligence and ethics: A comprehensive review of bias mitigation, transparency, and accountability in ai systems.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models.

Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. 2024. In-contextual gender bias suppression for large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1722–1742, St. Julian's, Malta. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Daniele Orner, Elizabeth Akinyi Ondula, Nick Mumero Mwangi, and Richa Goyal. 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*, pages 2408–2410.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Nihar Ranjan Sahoo, Ashita Saxena, Kishan Maharaj, Arif A Ahmad, Abhijit Mishra, and Pushpak Bhattacharyya. 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*, pages 73–79.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 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.

Hari Shrawgi, Prasanjit Rath, Tushar Singhal, and Sandipan Dandapat. 2024. Uncovering stereotypes in large language models: A task complexity-based approach.

In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1841–1857, St. Julian's, Malta. Association for Computational Linguistics.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. 2023. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023a. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023b. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting.

Eddie L Ungless, Amy Rafferty, Hrichika Nag, and Björn Ross. 2022. A robust bias mitigation procedure based on the stereotype content model. *arXiv preprint arXiv:2210.14552*.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao 'Kenneth' Huang, and Shomir Wilson. 2023. Nationality bias in text generation.

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

Danqing Wang and Lei Li. 2023b. Learning from mistakes via cooperative study assistant for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10667–10685, Singapore. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2203.11171*.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, 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.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

A Appendix

A.1 Benchmark Introduction

The BBQ dataset comprises a set of manually crafted question sets that cover nine social bias dimensions and two cross-dimensions within the context of American English. Each question is formatted as a multiple-choice QA problem. The three answer options for each question include two social groups and "*Cannot be determined*," representing bias, anti-bias, and neutral (unbiased) choices, respectively. And we measure the debiasing capability using the Bias Score defined in the BBQ dataset. The Bias Score ranges from -1 to 1: a positive value indicates a bias against the protected group, a negative value suggests anti-bias. A Bias Score of 0 indicates an ideal, unbiased model. As we aim to avoid both a biased and an anti-biased model, we calculate the absolute value of the Bias Score to measure the level of bias in LLMs. We also measure downstream task performance using the accuracy of Disambiguous questions. The questions are categorized as **Ambiguous** or **Disambiguous. Ambiguous** questions lack specific context within the text, making "Cannot be determined" the correct answer. In this case, if LLMs choose either social group for ambiguous questions, it indicates the presence of bias or anti-bias issues. **Disambiguous** questions provide an unambiguous context, with one of the two social group representations being the correct answer. Choosing the wrong group or "Cannot be determined" for Disambiguous questions can indicate a decline in downstream performance. Given that the BBQ dataset was not originally designed for multi-objective debiasing, we adapt our methodology by calculating the absolute value of the Bias Score in ambiguous contexts to assess bias levels accurately. An unbiased model achieves a Bias Score of 0, whereas a completely biased model reaches an absolute value of 1, encompassing both bias and anti-bias as manifestations of bias. For downstream task performance, we measure accuracy specifically in unambiguous contexts. We avoid using accuracy from ambiguous contexts, as it correlates closely with the Bias Score and does not adequately test the LLMs' capacity for logical reasoning. Since the BBQ dataset is not initially designed for multi-objective debiasing, we calculate the absolute value of the Bias score in ambiguous contexts to measure bias levels. An unbiased model has a Bias Score of 0, while an entirely biased model has an absolute value of 1, where we consider bias and anti-bias as bias. We also measure downstream task performance using accuracy in disambiguous contexts, and we don't use accuracy from ambiguous contexts because it is heavily related to Bias Score and choosing unknown couldn't test the LLMs' logical abilities.. A robust and impartial Language Model will remain unbiased toward any social group when the text is insufficient and provide a logical answer when the text is adequate, resulting in a Bias Score of zero and 100% accuracy for disambiguous questions. As previously mentioned, we measure the multi-objective target of LLMs on the BBQ dataset using the Bias Score defined in (Parrish et al., 2022). The Bias Score ranges from -1 to 1: a positive value indicates a bias against the protected group, a negative value suggests anti-bias (i.e., bias against the dominant group not typically subjected to prejudice), and a Bias Score of 0 indicates an ideal, unbiased model. As we aim to avoid both a biased and an anti-biased model, in our experiments, we define a new metric, the Bias Score, calculated as the absolute value of the Bias Score defined in (Parrish et al., 2022).

The StereoSet dataset measures stereotypes across four domains, including Gender, Profession, Race, and Religion . As the model must provide probabilities, we adapt it into a QA format. Each question has three answer options: stereotype, anti-stereotype, and unrelated. Each question is posed six times in different sequences to mitigate order influence. We use three metrics defined in (Nadeem et al., 2020): Stereotype Score (ss), Language Modeling Score (lms), and Idealized CAT Score (icat) to measure the bias level, the performance on downstream tasks, and the overall multi-objective performance.

A.2 Anti-Bias Prompts in Preliminary Experiments

Method	Prompts
ABP-0	Let's think about how to answer the question in a way that avoids bias or stereotyping.
ABP-1	Please ensure that your answer is unbiased and does not rely on stereotypes.
ABP-2	We should treat people from different socioeconomic statuses, sexual orientations, religions, races, physical appearances, nationalities, gender identities, disabilities, and ages equally. When we do not have sufficient information.
ABP-3	We should choose the unknown option, rather than making assumptions based on our stereotypes.
ABP-4	Really don't discriminate.

Table 6: Anti-Bias Prompts in Preliminary Experiments

A.3 Main Experiments Results

Method	**Bias Score**	**Delta(%)**	**Task Acc**	**Delta(%)**
Baseline	0.13845	-	0.86365	-
Pure CoT	0.13079	-5.5	0.80173	-7.2
Society of Mind	0.17183	+24.1	0.83465	-3.3
		using EquiSync methods		
Masking	**0.01696**	**-87.8**	0.81345	-5.8
Balancing	0.04281	-69.1	0.80502	-6.8

Table 7: **Llama-3-8B-Instruct on BBQ**

Method	**Bias Score**	**Delta(%)**	**Task Acc**	**Delta(%)**
Baseline	0.09362	-	0.83993	-
Pure CoT	0.08954	-4.7	0.87116	+3.7
Society of Mind	0.09139	-2.4	0.86982	+3.6
		using EquiSync methods		
Masking	0.01877	-80.0	0.89769	+6.9
Balancing	0.04545	-51.5	0.85292	+1.5

Table 8: **GPT-3.5-Turbo on BBQ**