UMAD: Enhancing LLM Debiasing via Multi-Agent Debate and Token-Level Bias Interpretation

Hanwen Gu, Jie Ma, Ying Qin, Ling Hu[†]
School of Information Science and Technology
Beijing Foreign Studies University, Beijing, China
guhanwen@bfsu.edu.cn, majie@bfsu.edu.cn, qinying@bfsu.edu.cn
huling@bfsu.edu.cn

Abstract

Textual data often contain biases that compromise fairness in AI systems, particularly in sensitive areas such as gender, race, and politics. While large language models (LLMs) have shown success across various tasks, they still face limitations due to inherent biases within the models and restrictive safety policies that hinder direct bias mitigation. To overcome these challenges, we propose UMAD (Unsupervised Multi-Agent Debate), a novel framework that leverages a Multi-Agent Debate mechanism alongside Best-Worst Scaling (BWS) to foster more effective discussions among LLMs, facilitating the identification of biases. By combining this with gradient-based interpretation techniques, UMAD extracts token-level bias insights, which are then integrated into models using in-context learning. This enhances the debiasing performance, as shown by our experiments across three bias categories—gender, religion, and politics—using five different LLMs. Our approach demonstrates significant improvements in metrics, with large models matching or even surpassing GPT-4 in Style Accuracy (STA). We release our code at: https://github.com/Couen/UMAD.git.

Warning: this paper contains content that may be offensive or upsetting.

1 Introduction

Bias in text, such as gender bias (Doughman and Khreich, 2023), political bias (Mou et al., 2023), and religious bias (Hu et al., 2022), presents significant challenges to the fairness and reliability of AI systems. With the rapid growth of Artificial Intelligence Generated Content (AIGC), these biases have become harder to detect, as both explicit and implicit biases can be reproduced (Wang et al., 2023; Felkner et al., 2023; Lee et al., 2023). This makes debiasing even more crucial, as biased AI-generated texts can propagate and amplify harmful stereotypes in real-world applications.

Debiasing textual data aims to mitigate these biases while preserving content quality. Prevalent debiasing methods primarily fall into two categories: data replacement and data generation. Data replacement techniques, such as counterfactual data augmentation (CDA) (Qian et al., 2022; Zayed et al., 2023) and selective masking (Thakur et al., 2023; Ghanbarzadeh et al., 2023), work by identifying and replacing biased words. However, these approaches often compromise sentence fluency and may introduce errors (Dale et al., 2021). On the other hand, data generation methods leverage auto-encoder sequence-to-sequence models to rewrite biased texts. These methods typically rely on large amounts of annotated data, which is expensive and difficult to obtain (Han et al., 2024). With the advent of large language models, many studies have explored their use in reducing biases, such as in models like ChatGPT (OpenAI, 2021), and LLaMA3 (MetaAI, 2024). However, this approach is further limited by the inherent biases of the LLMs themselves and the restrictive safety policies imposed on these models, which result in suboptimal debiasing performance (Nozza et al., 2022; Oba et al., 2024).

To address the limitations of existing methods, we propose the Unbiased Multi-Agent Debiasing (UMAD) framework, which focuses on extracting token-level bias insights using gradient-based interpretations and integrating these insights into LLMs through in-context learning. This allows for more

[†] Corresponding author.

targeted debiasing while preserving overall content quality. UMAD is fully unsupervised, eliminating the need for costly annotated data, and combines a Multi-Agent Debate mechanism with the Best-Worst Scaling (BWS) algorithm to enable collaborative bias annotation across multiple LLMs. By leveraging these techniques, UMAD enhances the accuracy of bias detection and improves the robustness of the debiasing process.

In summary, the key contributions of this paper are:

- We present the first unsupervised bias annotation framework that combines a Multi-Agent Debate mechanism with the BWS algorithm, addressing the scarcity of annotated bias data and improving the robustness of bias detection across diverse contexts and applications.
- We introduce a novel integration of gradient-based interpretations into the debiasing process, enabling the extraction of token-level bias information and guiding LLMs to focus on biased tokens during text generation.
- Our extensive experiments across five LLMs and three types of bias—politics, religion, and race—demonstrate that UMAD significantly improves bias mitigation, achieving a 13% improvement in the Unbias metric and a 20% increase in the Specific Task Accuracy (STA) metric. These results highlight UMAD's superiority over existing methods in both debiasing performance and content preservation.

2 Related Work

2.1 Text Debiasing

From the perspective of debiasing at token-level or sentence-level, existing research can be broadly divided into data replacement and data generation approaches.

The replacement approach is well researched due to its simplicity. It requires accurately identifying biased words and finding suitable replacements. Raza et al. (2023) suggested creating a list of biased words and locating them within texts. Hallinan et al. (2023) used the BERT model to identify biased named entities as biased keywords. Floto et al. (2023) introduced MARCO to identify biased words through comparative analysis by unbiased and biased experts. Meanwhile, embedding-based and model prediction-based methods have been widely used to find replacement words. The former uses word embeddings to find semantically similar words (Raza et al., 2023), while the latter generates similar words through model predictions (Floto et al., 2023).

The generation approach addresses debiasing at sentence-level, which leverages the text generation abilities of pre-trained models (Madanagopal and Caverlee, 2023; Raza et al., 2023), diffusion models (Floto et al., 2023), and LLMs (Mishra et al., 2024). They rewrote biased documents into unbiased ones. Typically, they employed an encoder-decoder architecture, where biased texts are encoded and then decoded to produce unbiased content (Madanagopal and Caverlee, 2023; Raza et al., 2023). LLMs exhibited impressive performance in data debiasing by rewriting biased documents (Pesaranghader et al., 2023). However, the hallucination of LLMs and their lack of transparent debiasing explanations, limit their applications.

For the perspective of relying on annotated data, existing debiasing research can be divided into supervised and unsupervised approaches. Data debiasing is often challenging due to the scarcity of annotated data (Raffel et al., 2020) and most debiasing methods are supervised (Raza et al., 2023; Floto et al., 2023; Madanagopal and Caverlee, 2023; Raza et al., 2023). Annotated data for one bias category is hard to extend into different bias categories. To address this issue, recent research has proposed to generate synthetic data to produce high-quality training data (Ouyang et al., 2022). Though synthetic data generated by LLMs exhibits generalizability across various bias categories, it remains a supervised approach, leaving unsupervised data debiasing largely unexplored.

2.2 Interpretable Technique

Interpretable technique can help understand a deep learning model's decision-makings, and are often categorized along two main aspects: *local* and *global* (Danilevsky et al., 2020). A *local* explanation

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1078-1094, Jinan, China, August 11-14, 2025.

provides information for the model's prediction on a specific input and a *global* explanation focuses on the overall predictive process of the model, independent of individual inputs.

Local interpretability approaches, such as LIME (Ribeiro et al., 2016), SHAP (Mosca et al., 2022), and gradient-based methods (Karlekar et al., 2018), have been demonstrated their capacities of analyzing models' debiasing decisions. For example, Devatine et al. (2023) utilized LIME to clarify the decision-making processes of pre-trained models for detecting political bias. Dhingra et al. (2023) employed SHAP to detect and mitigate gender bias of LLMs. Danilevsky et al. (2020) indicated that gradient-based methods are particularly effective in providing interpretability at the token-level feature. Therefore, we use the gradient-based approach as an interpretable technique to uncover insights into token-level biases.

2.3 Multi-Agent Debate

Multi-agent debate involves a group of LLMs engaged in argumentative interactions to make decisions or solve tasks. This approach was inspired by the suggestion that multiple LLMs can enhance each other's performance through debate and cooperation (Chan et al., 2023). Furthermore, research also demonstrated that debate is an effective method to address the inconsistency issue of single LLM (Xiong et al., 2023).

For example, ChatEval was a multi-agent referee team, which enables LLMs to act as human evaluators (Chan et al., 2023). LLMs have demonstrated annotation performance comparable to human annotators in certain domains (Ziems et al., 2024), but they may inherently carry biases from their training corpora, exhibiting a range of different biases (Feng et al., 2023). Inspired by these findings, we employ a multi-agent debate framework for bias annotation, enhancing annotation performance and mitigating the inherent biases of individual LLMs.

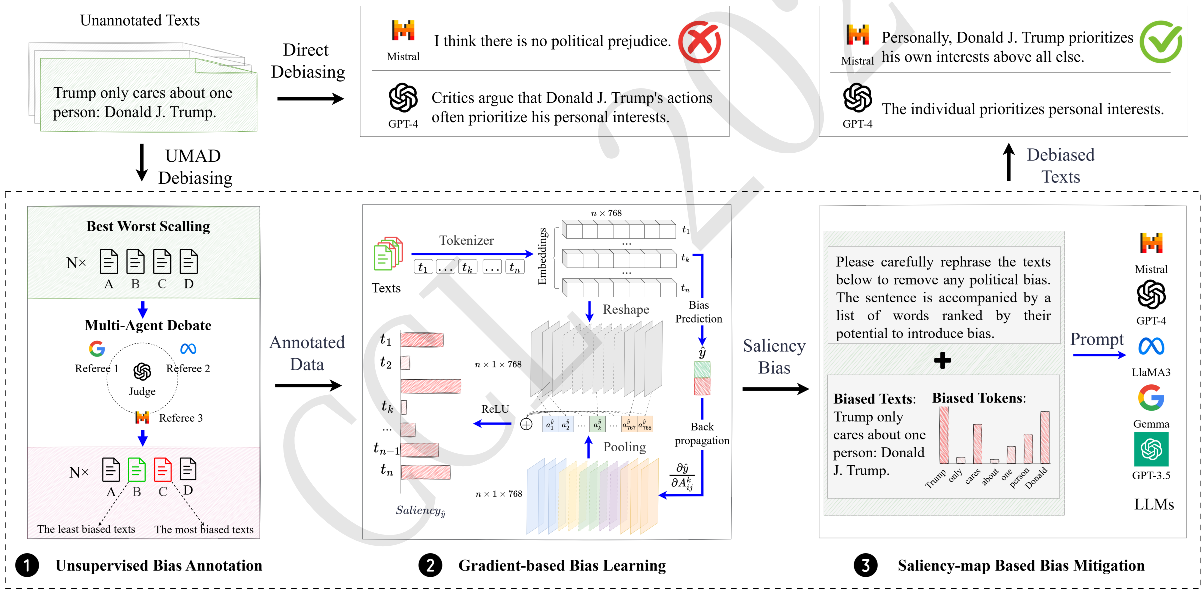


Figure 1: Illustration of UMAD framework, comprising three steps: ❶ Unsupervised Bias Annotation ❷ Gradient-based Bias Learning ❸ Saliency-map based Bias Mitigation.

3 UMAD Debiasing Framework

As illustrated in Figure 1, the proposed Unsupervised Multi-Agent Debate (UMAD) framework consists of three modules: ❶ Unsupervised Bias Annotation, ❷ Gradient-based Bias Learning, and ❸ Saliency-map based Bias Mitigation.

The first module in UMAD achieves unsupervised bias text annotation through four LLMs' collaborative debate, which serve as annotators in the Best-Worst Scaling (BWS) algorithm. The second module in UMAD employs a gradient-based interpretation technique to train a pre-trained model with the annotated biased data. The trained model then acts as a bias classifier and detector, capable of highlighting key

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1078-1094, Jinan, China, August 11-14, 2025.

biased words with saliency-map scores. Finally, the saliency-map based bias mitigation module aims to guide LLMs' debiasing process with token-level bias insights from these saliency-map scores.

3.1 Unsupervised Bias Annotation

3.1.1 Best Worst Scaling

Best-Worst Scaling, initially developed by Louviere and Woodworth (1991), is an efficient annotation framework using a comparative approach. It can generate highly reliable ratings by identifying the best and worst items according to a specific criterion, making it a valuable tool for various annotation tasks. For unsupervised bias annotation, it is challenging to annotate the data and quantitatively evaluate the results. By leveraging BWS, explicit comparisons can be made between biased data, thus achieving effective unsupervised bias annotation.

In BWS annotation procedure, data is firstly grouped into sets of n items, where annotators are asked to identify the most and least biased items in each group. In this work, n equals 4 as 4-tuples is particularly efficient in best-worst annotations (Hada et al., 2023).

For example, consider four items A, B, C, and D, where A is annotated as the most biased and D as the least biased. Then the following six comparisons can be derived: A>B, A>C, A>D, B>D, and C>D. By analyzing such comparisons across multiple 4-tuples, the real-valued bias score for each item can be computed. Specifically, the bias score of an item is calculated as the difference between the number of times it is selected as the most biased and the number of times it is selected as the least biased, normalized by the total number of times the item is included in comparisons (Orme, 2009).

3.1.2 Multi-Agent Debate

In multi-agent debate procedure, four LLMs are employed as annotators in the BWS bias ranking process: three as referees and one as the judge. These LLMs collaborate in debating and discussing to annotate bias for each item. As shown in Fig. 2 , the multi-agent debate includes four steps.

First, two LLMs engage in debate to determine the most and the least biased documents. Next, a judge evaluates whether the two LLMs' opinions are consistent. If they do reach an agreement, the judge will confirm their decisions. If not, a third referee will be consulted to provide an independent assessment to resolve the discrepancies. Finally, the judge reviews all opinions. If a consensus is reached, the judge will conclude the evaluation and label the most and the least bias documents. If not, the judge will declare the discussion invalid and discards the evaluation result, ensuring that only consistent annotations are used for further processing. Detailed description of the designed prompts can be found in the appendix (see Table 7 in Appendix B).

Table 1: Illustration of prompts designed to mitigate bias with saliency-map bias ranking information. Tokens in the sentece are highlighted with varying degree of red, indicating different levels of bias.

Role	Prompt
User	Please carefully rephrase the sentence below to remove any political bias. The sentence is accompanied by a list of words ranked by their potential to introduce bias. **Biased Tokens:** democracy, der, mafia, capitalism, failed, result, natural, united **Biased Text:** The United States is a failed democracy and a mafia state, the natural result of what happens when capitalism is deregulated.
Assis-tant	The United States has encountered challenges in its governance and economic regulation, leading to significant issues.

3.2 Gradient-based Bias Learning

After obtaining the annotated data, a gradient-based bias learning approach is utilized to extract token-level bias information to aid bias mitigation. In detail, we employ GradCAM (Selvaraju et al., 2016) to visualize the token-level bias features within each document. GradCAM demonstrates impressive potential in the field of Computer Vision to reveal the importance of feature maps through gradient analysis. Danilevsky

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1078-1094, Jinan, China, August 11-14, 2025.

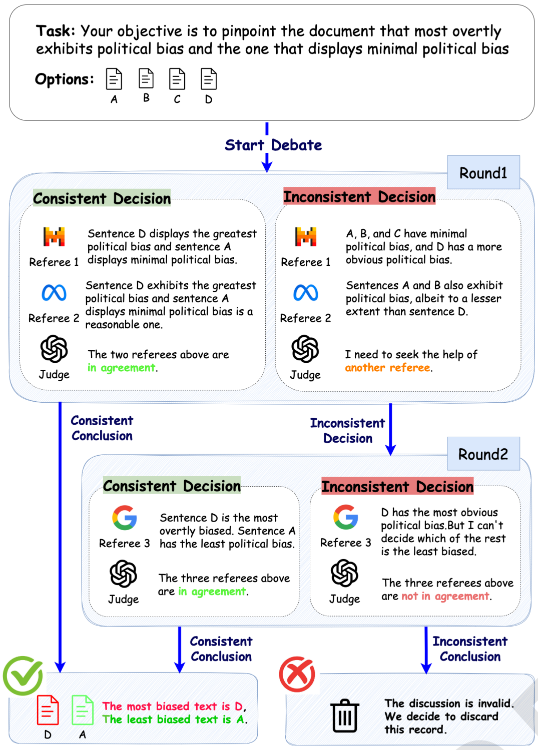


Figure 2: Illustration of the Multi-Agent Bias Debate. Three referees and a judge collaboratively debate to identify the most and least biased items for each group of four documents. Only the consistent annotations will be kept for further processing.

et al. (2020) indicate that gradient analysis can capture the importance of word/token-level features. Therefore, we use it to identify the token-level bias features.

Hence, We fine-tune a BERT model integrated with the GradCAM to obtain a bias classifier and token-level bias detector. Let $D = \{d_1, d_2, \ldots, d_m\}$ be a bias dataset composed of m documents, where d_j denotes the j-th document in D. Let $\hat{y_j}$ denote the predicted label of d_j by the fined-tuned BERT classifier, where $\hat{y_j}$ equals to 0 (for unbiased text) or 1 (for biased text).

The BERT model tokenizes d_j into a sequence of tokens $\{t_1, t_2, \ldots, t_n\}$, where n is the number of tokens in d_j. Given tokens as input, we can extract the last layer of BERT to obtain the embedding representation of the document, denoted as $E \in \mathbb{R}^{n\times 768}$. Then we reshape E to have $E_{\text{reshaped}} \in \mathbb{R}^{n\times 1\times 768}$, where $E_{\text{reshaped}} = \{A^1, A^2, \ldots, A^{768}\}$ and A^k is the k-th feature map, $k \in [1, \ldots, 768]$.

In order to obtain a class-discriminative localization saliency map $\text{Saliency}_{\hat{y_j}}$ for the predicted label $\hat{y_j}$, we compute the gradient weights $\alpha_k^{\hat{y_j}}$, with respect to $A^k \in \mathbb{R}^{n\times 1}$ by performing global average pooling on the gradients across the spatial dimensions, as follows:

$$\alpha_k^{\hat{y_j}} = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial \hat{y_j}}{\partial A_{ij}^k}}_{\text{gradients via backprop}} \tag{1}$$

where $\frac{\partial \hat{y_j}}{\partial A_{ij}^k}$ represents the gradient of the k-th feature map.

Then each gradient weight $\alpha_k^{\hat{y_j}}$ is combined with the corresponding feature map A^k using weight sum calculation, followed by the ReLU activation function for generating the saliency map $\text{Saliency}_{\hat{y_j}}$:

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1078-1094, Jinan, China, August 11-14, 2025.

$$\text{Saliency}_{\hat{y_j}} = \text{ReLU}\underbrace{\left(\sum_{k=1}^{768} \alpha_k^{\hat{y_j}} A^k\right)}_{\text{linear combination}} \tag{2}$$

Finally, the documents classified as biased texts by BERT classifier, we use $\text{Saliency}_{\hat{y_j}}$ to represent the bias degree of each token in document d_j.

3.3 Saliency-map based Bias Mitigation

Given $\text{Saliency}_{\hat{y_j}}$ of document d_j, the saliency-map based bias mitigation aims to integrate these bias degree information into LLMs' debiasing prompts.

We first remove the stop words and then define a sorting function SortTokens to sort the tokens in d_j according to their bias degree in $\text{Saliency}_{\hat{y_j}}$, calculated as:

$$T_j^{\text{sorted}} = \text{SortTokens}(d_j|\text{Saliency}_{\hat{y_j}}) \tag{3}$$

Finally, LLMs will debias on d_j to obtain $\hat{d}_j$, implemented as:

$$\hat{d}_j = \text{DebiasLLM}(\text{Prompt}|T^j_{\text{sorted}}, d_j) \tag{4}$$

Table 1 illustrates the detailed description of the designed Prompt for debiasing with saliency-map bias ranking information.

4 Experiments

Category	Metric	Baselines					UMAD					Avg Inc.
		Gemma	Mistral	LLaMA3	GPT-3.5	GPT-4	Gemma	Mistral	LLaMA3	GPT-3.5	GPT-4	
Politics	UnBias	0.00	0.01	0.02	0.89	0.41	0.58 ↑	0.58 ↑	0.58 ↑	0.58 ↓	0.58 ↑	**31%** ↑
	STA	0.45	0.14	0.56	0.20	0.82	0.36 ↓	0.73 ↑	0.66 ↑	0.65 ↑	0.70 ↓	**18%** ↑
	Un+STA	0.29	0.40	0.39	0.57	0.62	0.47 ↑	0.66 ↑	0.62 ↑	0.62 ↑	0.64 ↑	**14%** ↑
	F1	0.65	0.66	0.65	0.41	0.74	0.73 ↑	0.73 ↑	0.73 ↑	0.73 ↑	0.73 ↓	**10%** ↑
	SIM	0.78	0.58	0.70	0.68	0.63	0.50 ↓	0.49 ↓	0.56 ↓	0.59 ↓	0.57 ↓	-13% ↓
Religion	UnBias	0.03	0.49	0.31	0.98	0.41	0.66 ↑	0.66 ↑	0.66 ↑	0.66 ↓	0.66 ↑	**21%** ↑
	STA	0.22	0.21	0.24	0.01	0.36	0.41 ↑	0.42 ↑	0.56 ↑	0.41 ↑	0.53 ↑	**25%** ↑
	Un+STA	0.14	0.32	0.27	0.40	0.38	0.52 ↑	0.52 ↑	0.60 ↑	0.51 ↑	0.58 ↑	**24%** ↑
	F1	0.73	0.74	0.74	0.03	0.83	0.72 ↓	0.72 ↓	0.72 ↓	0.72 ↑	0.72 ↓	**10%** ↑
	SIM	0.62	0.53	0.53	0.51	0.48	0.48 ↓	0.42 ↓	0.39 ↓	0.51 –	0.48 –	-7% ↓
Race	UnBias	0.23	0.74	0.38	0.93	0.30	0.40 ↑	0.40 ↓	0.40 ↑	0.40 ↓	0.40 ↑	-11% ↓
	STA	0.54	0.10	0.18	0.03	0.45	0.47 ↓	0.36 ↑	0.54 ↑	0.40 ↑	0.48 ↑	**19%** ↑
	Un+STA	0.44	0.31	0.25	0.33	0.40	0.45 ↑	0.38 ↑	0.49 ↑	0.40 ↑	0.46 ↑	**9%** ↑
	F1	0.78	0.59	0.71	0.25	0.80	0.74 ↓	0.74 ↑	0.74 ↑	0.74 ↑	0.74 ↓	**11%** ↑
	SIM	0.41	0.48	0.39	0.53	0.76	0.47 ↑	0.46 ↓	0.32 ↓	0.54 ↑	0.56 ↓	-4% ↓

Table 2: Debiasing performance comparison across bias categories of Politics, Religion, and Race in five LLMs, as well as their improvements by UMAD.

4.1 Experimental Setup

4.1.1 Datasets

We evaluate our framework across three bias categories: Politics, Race and Religion, sourced from datasets of MBIC (Spinde et al., 2021) and ToxicBias (Sahoo et al., 2022).

Specifically, MBIC is a political bias dataset sourced from various media outlets. It comprises 3,700 sentences, with 1,863 unbiased documents and 1,860 biased sentences. ToxicBias (Sahoo et al., 2022) is an unintended bias dataset from a Kaggle competition, encompassing five types of biases: "Politics",

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1078-1094, Jinan, China, August 11-14, 2025.

"Race", "Religion", "Gender" and "LGBTQ". For our study, we focus specifically on religious and racial biases, which are the two most prevalent types in the dataset. Specifically, it includes 1,575 instances of religious bias, 2,203 instances of racial bias, and 1,084 unbiased instances.

4.1.2 Baselines and Metrics

We use five LLMs as baselines to implement direct debiasing, including Gemma-7B-it from Gemma Team (Gemma Team et al., 2024) (denoted as Gemma), Mistral-7B-Instruct from (Jiang et al., 2023) (denoted as Mistral), LLaMA3-8B-Instruct from Meta (MetaAI, 2024) (denoted as LLaMA3), and two LLMs from OpenAI: gpt-3.5-turbo-0125 (OpenAI, 2021) (denoted as ChatGPT) and gpt-4-0125-preview (OpenAI, 2024) (denoted as GPT-4). During the multi-agent debate process, we employed three smaller-scale models: Gemma, Mistral, and LLaMA3 as referees, and ChatGPT as the judge for scoring. In addition, we compared three traditional models that can be used for text debiasing: CondBERT (Dale et al., 2021), ParaGeDi (Dale et al., 2021), and Marco (Hallinan et al., 2023), which are traditional methods leveraging pre-trained language models.

In our text debiasing task, we utilize five evaluation metrics: F1, Unbias, Style Accuracy (STA) (Logacheva et al., 2022), Un+STA, and Content Preservation (SIM) (Pour et al., 2023). Notably, Unbias and Un+STA are our new metrics designed to comprehensively assess the model's debiasing effectiveness while avoiding redundant debias on unbiased texts, addressing limitations of previous metrics. Detailed calculation steps for the metrics can be found in the appendix (see Appendix B).

4.2 Experimental Results

4.2.1 Debias Analysis

We compare the debiasing performance of UMAD against the baseline models on Politics, Religion, and Race bias. The results are reported in Table 2. Our framework almost outperforms all the baseline models across all metrics except SIM scores, indicating the framework can improve bias identification even on finer-grained token-level. Besides, lower SIM scores for UMAD are reasonable since the biased documents require substantial rewriting to become unbiased, and they tend to be less similar to the original biased text used as input.

Interestingly, we observe that our method shows greater improvements on smaller-scale models like LLaMA3 and Mistral. With ranked biased token-level information, LLaMA3 exhibits notable improvements of 23%, 33%, and 24% in Un+STA scores across three datasets, while Mistral demonstrates enhancements of 26%, 20%, and 7% in corresponding datasets. In models where performance are already strong, such as GPT-4, the UMAD maintains or even further improves debiasing performance, which further demonstrates the effectiveness of our proposed framework.

Additionally, models perform unbalance on different bias categories. Both the baseline models and the UMAD framework perform better in mitigating political bias. This could be attributed to political bias being more straightforward and less sensitive in datasets. For racial and religious biases, baseline models show lower performance on the STA score, suggesting that LLMs struggle to accurately identify key biased tokens during direct debiasing. Note that, our method with token-level bias information can improve STA scores, enabling more precise debiasing.

Furthermore, we also compare the debiasing performance of LLM-based methods against traditional methods across bias categories of Politics, Religion, and Race. Table 3 shows the result, where LLMs and UMAD represent the average performance of 5 baseline LLMs. LLM-based methods significantly outperform traditional methods on UnBias, STA, and Un+STA, especially with UMAD. Notably, traditional methods achieve best scores on F1 and SIM while get 0 on UnBias, indicating that they tend to classify all the texts as biased and have insufficient ability to identify bias accurately.

4.2.2 Ablation Experiments

We conduct two sets of additional ablation studies. The first aims to validate the effectiveness of Best Worst Scaling in identifying bias, while the second seeks to investigate the impact of bias levels in training data on learning bias.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1078-1094, Jinan, China, August 11-14, 2025.

Category	Metric	CondBERT	ParaGeDi	Marco	LLMs	UMAD
Politics	UnBias	0.00	0.00	0.54	0.26	**0.58**
	STA	0.23	0.07	0.10	0.43	**0.62**
	Un+STA	0.11	0.03	0.32	0.45	**0.60**
	F1	0.66	0.66	0.56	0.62	**0.73**
	SIM	**0.95**	0.39	0.91	0.67	0.54
Religion	UnBias	0.00	0.00	0.28	0.44	**0.66**
	STA	0.08	**0.46**	0.15	0.20	**0.46**
	Un+STA	0.05	0.27	0.20	0.30	**0.54**
	F1	**0.74**	**0.74**	0.67	0.61	0.72
	SIM	**0.93**	0.55	0.88	0.53	0.45
Race	UnBias	0.00	0.00	0.28	**0.51**	0.40
	STA	0.10	0.27	0.12	0.26	**0.45**
	Un+STA	0.07	0.18	0.17	0.34	**0.43**
	F1	**0.80**	**0.80**	0.67	0.62	0.74
	SIM	**0.92**	0.53	0.87	0.51	0.47

Table 3: Debiasing comparisons between traditional methods and LLMs across biases. Bold values indicate the highest performance.

To understand the necessity of BWS, we conduct first ablation experiment on a political dataset with two schemes: without BWS(w/o BWS) and BWS, where 'w/o BWS' denotes referees directly assess bias for each text. Table 4 shows the result of debiasing performance with and without BWS on three models. Compared to models without BWS, models with BWS achieved higher scores on four of five metrics, indicating that BWS can significantly improve the detection of bias within texts.

Method	Model	Unbias	STA	Un+STA	F1	SIM
w/o BWS	Gemma	**0.90**	0.15	0.53	0.53	0.48
	LLaMA3	**0.90**	0.32	0.61	0.53	0.33
	GPT-3.5	**0.90**	0.31	0.61	0.53	0.56
BWS	Gemma	0.58	0.36	0.47	**0.73**	0.48
	LLaMA3	0.58	**0.66**	**0.62**	**0.73**	0.32
	GPT-3.5	0.58	0.65	**0.62**	**0.73**	**0.58**

Table 4: Ablation study results on the Political debiasing with and without BWS algorithm. Bold values indicate the highest performance.

To investigate the impact of bias levels in training data on learning bias, we conduct second ablation experiment on a race dataset with different training data. The degree of bias is determined with BWS scores and can be categorized into four levels: extreme ([-1,-0.8) ∪ (0.8,1]), moderate ([-1,-0.5) ∪ (0.5,1]), mild ([-1,-0.3) ∪ (0.3,1]), and full range ([-1,0) ∪ (0,1]). The results are shown in Table 5, indicating two interesting findings. First, unbiased training data has minimal effect on model's performance. Compared to the full range, utilizing data from moderate range achieves the comparable performance with the highest F1 and the optimal Un+STA. It because the discarded data outside the range contain mostly unbiased text. Second, besides bias level, training data quantity is also crucial. The performance by using part of the annotation data from extreme range drops, indicating that the debiasing ability of UMAD also decreases when training samples are insufficient.

4.2.3 Interpretable Case Study

In this section, we carry out a case study to validate the effectiveness of UMAD's interpretablity. By visualizing the top 120 topic words in LLaMA3's successfully and unsuccessfully debiased texts in the directly debiasing setting, we can gain some insights about LLMs' debiasing preference and disregard. The visualization is shown in Figure 3.

Firstly, as depicted in Figure 3a, LLaMA3 naturally focus on terms like "Mainstream", "Student", and "Transgender", indicating LLaMA3's proficiency in handling biases related to education and social issues. Secondly, as highlighted in Figure 3b, LLaMA3 struggled with the terms like "Trump", "Republicans"

Score	Model	Unbias	STA	Un+STA	F1	SIM
Extreme	Gemma	0.70	0.29	0.42	0.60	0.48
	LLaMA3	0.70	0.34	0.46	0.60	0.36
	GPT-3.5	0.70	0.23	0.39	0.60	0.52
Moderate	Gemma	0.40	**0.47**	0.45	**0.74**	0.47
	LLaMA3	0.40	**0.54**	0.49	**0.74**	0.37
	GPT-3.5	0.40	**0.40**	0.40	**0.74**	0.54
Mild	Gemma	**0.81**	0.24	0.43	0.58	**0.50**
	LLaMA3	**0.81**	0.32	0.48	0.58	0.38
	GPT-3.5	**0.81**	0.17	0.38	0.58	**0.57**
Full Range	Gemma	0.66	0.38	**0.47**	0.69	0.48
	LLaMA3	0.66	0.44	**0.51**	0.69	**0.38**
	GPT-3.5	0.66	0.31	**0.43**	0.69	0.53

Table 5: Ablation study results on Race debiasing on different ranges of BWS scores. Bold values indicate the highest performance.

Category	Bias Keywords				
Politics	rights	trump	house	republican	democrats
	gun	black	police	president	american
	left	party	corona	donald	administration
	law	racist	media	radical	democratic
	bad	lives	white	social	government
	tax	wing	matter	health	movement
Religion	allah	islam	islamic	terrorist	religion
	jews	kill	muslims	canada	women
	law	gay	radical	islamist	terrorism
	time	anti	death	countries	catholic
	mass	isis	religions	terrorists	american
	god	white	racist	innocent	christians
Race	guy	kill	white	racist	folks
	bad	anti	black	illegal	america
	cops	race	police	person	mexican
	shot	real	lives	males	women
	live	hate	crime	matter	violence
	cop	life	trash	racism	privilege

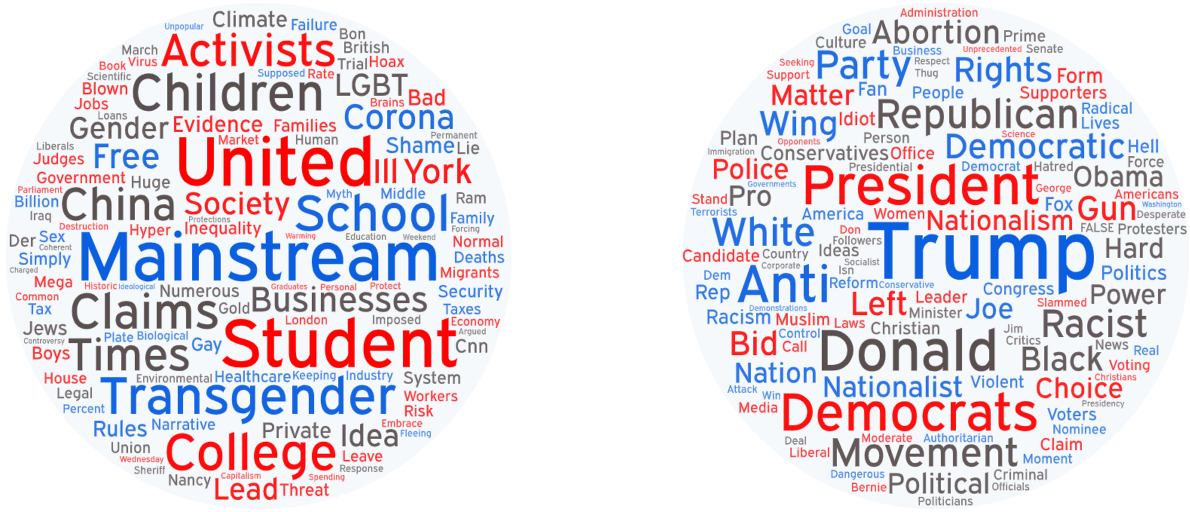
Table 6: Top 30 biased keywords with the highest scores in each category, obtained from UMAD.

and "Democrats", which are significantly biased in political contexts, revealing inherent political bias in LLaMA3.

Furthermore, we conduct a statistical analysis across political, religious, and racial datasets. Using our proposed bias learning method, we extract the top 30 keywords most associated with bias, along with their saliency scores, as detailed in Table 6. Specifically, in the Politics category, terms such as "democrat," "trump", and "republican" are identified as highly relevant to bias, while the Religious category highlights words like "muslims", "islam", and "jews". In the Racial category, terms such as "white", "black", and "racist" show strong associations with bias. These words are identified as closely related to bias through our gradient-based interpretability method, underscoring the effectiveness of our approach in revealing token-level bias relevance.

Notably, our method successfully identifies bias-relevant words in the Politics domain, where LLaMA3 faces challenges, providing valuable token-level insights to LLMs and demonstrating the efficacy of UMAD in bias detection and interpretation.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1078-1094, Jinan, China, August 11-14, 2025.



(a) LLaMA3's debiasing preference in political bias.

(b) LLaMA3's debiasing disregard in political bias.

Figure 3: Word clouds of the LLaMA3's preference and disregard in mitigating political bias. (a) keywords that LLaMA3 effectively debias. (b) keywords that LLaMA3 struggle to debias.

5 Conclusion

In this paper, we propose UMAD, an Unsupervised Multi-Agent Debate framework. UMAD begins with a collaborative multi-agent debate using the BWS algorithm for unsupervised bias annotation. We then employ an interpretable technique to provide token-level bias insights to LLMs. Experimental results demonstrate that our method significantly enhances LLMs' debiasing performance, particularly benefiting smaller-scale models like LLaMA3. This approach not only improves debiasing capabilities but also offers valuable insights into token-level biases, contributing to transparent and reliable bias mitigation strategies. Future work will focus on refining these techniques and exploring their applicability across diverse contexts and languages.

6 Acknowledgements

This work was supported by the National Social Science Foundation (No. 22BYY055). The authors would like to express their sincere gratitude to Yuemei Xu and Lin Gui for their valuable guidance and support throughout this research. We also thank Zihan Qiu for her helpful discussions and assistance. Finally, we appreciate the anonymous reviewers for their constructive comments and suggestions that helped improve this paper.

7 Limitations

Our UMAD framework shows promising results but also presents certain limitations. The current study focuses on overall debiasing performance, without deeply exploring how the multi-agent debate mitigates individual model biases. While the debate reduces individual model biases, collective biases may still affect results. Future work will investigate the debate mechanism's impact on individual bias reduction, along with refinements such as fine-tuning with unbiased datasets or post-processing techniques to further mitigate LLM biases.

Additionally, the datasets used for evaluation are limited. Although UMAD has proven effective across three bias categories and five LLMs, broader testing on diverse datasets is needed to confirm the generalizability of the results. Since this study is fully unsupervised and emphasizes debiasing performance, future work will also provide a more detailed analysis of the debiasing process.

References

Shaina Raza, Syed Raza Bashir, Sneha, and Urooj Qamar. 2023. Addressing Biases in the Texts Using an End-to-End Pipeline Approach. In *Proceedings of the International Workshop on Algorithmic Bias in Search and Recommendation*, pages 100–107. Springer.

Skyler Hallinan, Alisa Liu, Yejin Choi, and Maarten Sap. 2023. Detoxifying Text with MaRCo: Controllable Revision with Experts and Anti-Experts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–242, Toronto, Canada. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically Neutralizing Subjective Bias in Text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):480–489.

Griffin Floto, Mohammad Mahdi Abdollah Pour, Parsa Farinneya, Zhenwei Tang, Ali Pesaranghader, Manasa Bharadwaj, and Scott Sanner. 2023. DiffuDetox: A Mixed Diffusion Model for Text Detoxification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7566–7574, Toronto, Canada. Association for Computational Linguistics.

Karthic Madanagopal and James Caverlee. 2023. Reinforced Sequence Training based Subjective Bias Correction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2585–2598, Dubrovnik, Croatia. Association for Computational Linguistics.

Jad Doughman and Wael Khreich. 2023. Gender Bias in Text: Labeled Datasets and Lexicons. arXiv preprint arXiv:2201.08675.

Masahiro Kaneko and Danushka Bollegala. 2021. Dictionary-based Debiasing of Pre-trained Word Embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 212–223, Online. Association for Computational Linguistics.

Shaina Raza, Chen Ding, and Deval Pandya. 2023. Mitigating Bias in Conversations: A Hate Speech Classifier and Debiaser with Prompts. arXiv preprint arXiv:2307.10213.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with Parallel Data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Mohammad Mahdi Abdollah Pour, Parsa Farinneya, Manasa Bharadwaj, Nikhil Verma, Ali Pesaranghader, and Scott Sanner. 2023. COUNT: COntastive UNlikelihood Text Style Transfer for Text Detoxification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8658–8666, Singapore. Association for Computational Linguistics.

Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil Rephrases Of Toxic Texts With Self-Supervised Transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1442–1461, Online. Association for Computational Linguistics.

Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting Unintended Social Bias in Toxic Language Datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 132–143, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shenggang Hu, Jabir Alshehabi Al-Ani, Karen D. Hughes, Nicole Denier, Alla Konnikov, Lei Ding, Jinhan Xie, Yang Hu, Monideepa Tarafdar, Bei Jiang, Linglong Kong, and Hongsheng Dai. 2022. Balancing Gender Bias in Job Advertisements With Text-Level Bias Mitigation. *Frontiers in Big Data*, 5. https://www.frontiersin.org/articles/c14.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Deepa Muralidhar. 2021. Examining Religion Bias in AI Text Generators. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*, pages 273–274, Virtual Event, USA. Association for Computing Machinery.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1078-1094, Jinan, China, August 11-14, 2025.

Xinyi Mou, Zhongyu Wei, Qi Zhang, and Xuanjing Huang. 2023. UPPAM: A Unified Pre-training Architecture for Political Actor Modeling based on Language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11996–12012, Toronto, Canada. Association for Computational Linguistics.

Julie Jiang, Xiang Ren, and Emilio Ferrara. 2023. Retweet-BERT: Political Leaning Detection Using Language Features and Information Diffusion on Social Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):459–469.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. POLITICS: Pretraining with Same-story Article Comparison for Ideology Prediction and Stance Detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States. Association for Computational Linguistics.

Jordan J. Louviere and George G. Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Working paper.

Rishav Hada, Agrima Seth, Harshita Diddee, and Kalika Bali. 2023. "Fifty Shades of Bias": Normative Ratings of Gender Bias in GPT Generated English Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*. https://api.semanticscholar.org/CorpusID:264490615.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining Inter-Consistency of Large Language Models Collaboration: An In-depth Analysis via Debate. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*. https://api.semanticscholar.org/CorpusID:258832565.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shan Zhang, Jie Fu, and Zhiyuan Liu. 2023. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. *arXiv preprint arXiv:2308.07201*. https://api.semanticscholar.org/CorpusID:260887105.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Terry Flynn and Anthony A. J. Marley. 2014. Best-worst scaling: Theory and methods. https://api.semanticscholar.org/CorpusID:51732624.

Rajeev Verma, Rajarshi Roychoudhury, and Tirthankar Ghosal. 2022. The Lack of Theory is Painful: Modeling Harshness in Peer Review Comments. In *Proceedings of the AACL 2022*. https://api.semanticscholar.org/CorpusID:253762075.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128:336–359.

Yangyang Shu, Anton van den Hengel, and Lingqiao Liu. 2023. Learning Common Rationale to Improve Self-Supervised Representation for Fine-Grained Visual Recognition Problems. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11392–11401.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural Media Bias Detection Using Distant Supervision with BABE – Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and others. 2024. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

MetaAI. 2024. Introducing Meta Llama 3: The Most Capable Openly Available LLM to Date. https://ai.meta.com/blog/meta-llama-3/. Accessed: 2024-05-10.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1078-1094, Jinan, China, August 11-14, 2025.

OpenAI. 2021. chatgpt: A Large-Scale Generative Model for Open-Domain Chat. https://github.com/openai/gpt-3.

OpenAI. 2024. GPT-4 Research. https://openai.com/index/gpt-4-research/. Accessed: 2024-05-10.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2024. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. *arXiv preprint arXiv:2010.00711*.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A Diagnostic Study of Explainability Techniques for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic Interpretability of Machine Learning. *arXiv preprint arXiv:1606.05386*.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-Based Explanation Methods: A Review for NLP Interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4593–4603, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nicolas Devatine, Philippe Muller, and Chloé Braud. 2023. An Integrated Approach for Political Bias Prediction and Explanation Based on Discursive Structure. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11196–11211, Toronto, Canada. Association for Computational Linguistics.

Harnoor Dhingra, Preetiha Jayashanker, Sayali Moghe, and Emma Strubell. 2023. Queer People are People First: Deconstructing Sexual Identity Stereotypes in Large Language Models. *arXiv preprint arXiv:2307.00101*.

Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018. Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 701–707, New Orleans, Louisiana. Association for Computational Linguistics.

Mohammad N. S. Jahromi, Satya M. Muddamsetty, Asta Sofie Stage Jarlner, Anna Murphy Høgenhaug, Thomas Gammeltoft-Hansen, and Thomas B. Moeslund. 2024. SIDU-TXT: An XAI Algorithm for NLP with a Holistic Assessment Approach. *arXiv preprint arXiv:2402.03043*.

Philip Mavrepis, Georgios Makridis, Georgios Fatouros, Vasileios Koukos, Maria Margarita Separdani, and Dimosthenis Kyriazis. 2024. XAI for All: Can Large Language Models Simplify Explainable AI? *arXiv preprint arXiv:2401.13110*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large Language Model based Multi-Agents: A Survey of Progress and Challenges. *arXiv preprint arXiv:2402.01680*.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining Inter-Consistency of Large Language Models Collaboration: An In-depth Analysis via Debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7572–7590, Singapore. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. *arXiv preprint arXiv:2305.19118*.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1078-1094, Jinan, China, August 11-14, 2025.

Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. Bias of AI-Generated Content: An Examination of News Produced by Large Language Models. *arXiv preprint arXiv:2309.09825*.

Lijing Wang, Yingya Li, Timothy Miller, Steven Bethard, and Guergana Savova. 2023. Two-Stage Fine-Tuning for Improved Bias and Variance for Large Pretrained Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15746–15761, Toronto, Canada. Association for Computational Linguistics.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140, Toronto, Canada. Association for Computational Linguistics.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-woo Ha. 2023. KoSBI: A Dataset for Mitigating Social Bias Risks Towards Safer Large Language Model Applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 208–224, Toronto, Canada. Association for Computational Linguistics.

Bryan Orme. 2009. Maxdiff Analysis: Simple Counting, Individual-Level Logit, and HB. *Sawtooth Software*.

Shaz Furniturewala, Surgan Jandial, Abhinav Java, Pragyan Banerjee, Simra Shahid, Sumit Bhatia, and Kokil Jaidka. 2024. Thinking Fair and Slow: On the Efficacy of Structured Prompts for Debiasing Language Models. *arXiv preprint arXiv:2405.10431*.

Jiao Ou, Jinchao Zhang, Yang Feng, and Jie Zhou. 2022. Counterfactual Data Augmentation via Perspective Transition for Open-Domain Dialogues. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1635–1648, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation Augmentation for Fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abdelrahman Zayed, Prasanna Parthasarathi, Gonçalo Mordido, Hamid Palangi, Samira Shabanian, and Sarath Chandar. 2023. Deep Learning on a Healthy Data Diet: Finding Important Examples for Fairness. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence, the Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, and the Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23)*, article 1637, 9 pages. AAAI Press.

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. 2023. Language Models Get a Gender Makeover: Mitigating Gender Bias with Few-Shot Data Interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–351, Toronto, Canada. Association for Computational Linguistics.

Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. 2023. Gender-tuning: Empowering Fine-tuning for Debiasing Pre-trained Language Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5448–5458, Toronto, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint arXiv:1910.10683*.

Ali Pesaranghader, Nikhil Verma, and Manasa Bharadwaj. 2023. GPT-DETOX: An In-Context Learning-Based Paraphraser for Text Detoxification. In *Proceedings of the 2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1528–1534. IEEE.

Ashish Mishra, Gyanaranjan Nayak, Suparna Bhattacharya, Tarun Kumar, Arpit Shah, and Martin Foltin. 2024. LLM-Guided Counterfactual Data Generation for Fairer AI. In *Companion Proceedings of the ACM on Web Conference 2024 (WWW '24)*, pages 1538–1545, Singapore. Association for Computing Machinery.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1078-1094, Jinan, China, August 11-14, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text Detoxification using Large Pre-trained Neural Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pengrui Han, Rafal Kocielnik, Adhithya Saravanan, Roy Jiang, Or Sharir, and Anima Anandkumar. 2024. ChatGPT-Based Data Augmentation for Improved Parameter-Efficient Debiasing of LLMs. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 73–105, St. Julian's, Malta. Association for Computational Linguistics.

Zhihong Sun, Chen Lyu, Bolun Li, Yao Wan, Hongyu Zhang, Ge Li, and Zhi Jin. 2024. Enhancing Code Generation Performance of Smaller Models by Distilling the Reasoning Ability of LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5878–5895, Torino, Italia. ELRA and ICCL.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Ze-Feng Gao, Kun Zhou, Peiyu Liu, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Small Pre-trained Language Models Can be Fine-tuned as Large Models via Over-Parameterization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3819–3834, Toronto, Canada. Association for Computational Linguistics.

Jinyuan Wang, Junlong Li, and Hai Zhao. 2023. Self-prompted Chain-of-Thought on Large Language Models for Open-domain Multi-hop Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2717–2731, Singapore. Association for Computational Linguistics.

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.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626. IEEE.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1):237–291. MIT Press.

Yifei Li, Lyle Ungar, and João Sedoc. 2023. Conceptor-Aided Debiasing of Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10703–10727, Singapore. Association for Computational Linguistics.

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.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Pipelines for Social Bias Testing of Large Language Models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 68–74, virtual+Dublin. Association for Computational Linguistics.

Muhao Chen, Chaowei Xiao, Huan Sun, Lei Li, Leon Derczynski, Anima Anandkumar, and Fei Wang. 2024. Combating Security and Privacy Issues in the Era of Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 8–18, Mexico City, Mexico. Association for Computational Linguistics.

Deonna M. Owens, Ryan A. Rossi, Sungchul Kim, Tong Yu, Franck Dernoncourt, Xiang Chen, Ruiyi Zhang, Jiuxiang Gu, Hanieh Deilamsalehy, and Nedim Lipka. 2024. A Multi-LLM Debiasing Framework. *arXiv preprint arXiv:2409.13884*. https://arxiv.org/abs/2409.13884

Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. 2024. Steering LLMs Towards Unbiased Responses: A Causality-Guided Debiasing Framework. *arXiv preprint arXiv:2403.08743*. https://arxiv.org/abs/2403.08743

Ipek Baris Schlicht, Defne Altiok, Maryanne Taouk, and Lucie Flek. 2024. Pitfalls of Conversational LLMs on News Debiasing. In *Proceedings of the First Workshop on Language-driven Deliberation Technology (DELITE) @ LREC-COLING 2024*, pages 33–38, Torino, Italia. ELRA and ICCL.

A Multi-Agent Debias Prompt

We present the multi-agent debias prompt used by three referees and one judge in the unsupervised bias annotation, as shown in Table 7.

Setting	Prompt
Referee	**System:** As a referee, your task involves collaborating with fellow referees to analyze four sentences. Your objective is to pinpoint the sentence that most overtly exhibits gender bias and the one that displays minimal gender bias. **User:** Four documents are as follows: $\{D_{\text{tuple}}\}$.
Judge (Round 1)	**System:** You are a judge, and you need to determine the final answer to the question by reading the discussion records of several referees. **User:** $\{D_{\text{tuple}}\}$. Referee 1: {Response}. Referee 2: {Response}.
Judge (Round 2)	**System:** You are a judge, and you need to determine the final answer to the question as much as possible by reading the discussion records of several referees. **User:** $\{D_{\text{tuple}}\}$. Referee 1: {Response}. Referee 2: {Response}. Referee 3: {Response}.

Table 7: Designed prompts used by referees and judges during the multi-agent debate process for unsupervised bias annotation, where $\{D_{\text{tuple}}\}$ represents the set of four sentences. {Response} represents the opinions provided by each referee.

B Evaluation Metrics - Detailed Explanations

Unbias: The Unbias metric measures the proportion of unbiased samples identified by the model, defined as:

$$\text{Unbias} = \frac{N_{\text{unbiased}}}{N_{\text{total\_unbiased}}} \tag{5}$$

where N_{unbiased} is the number of unbiased samples identified by the model, and $N_{\text{total\_unbiased}}$ is the total number of unbiased samples in the dataset.

Style Accuracy (STA): The proportion of unbiased samples after debiasing that were originally biased in the dataset, defined as: in bias identification, calculated as:

$$\text{STA} = \frac{N_{\text{unbiased\_after}}}{N_{\text{total\_biased}}} \tag{6}$$

where $N_{\text{unbiased\_after}}$ is the number of samples that are unbiased after debiasing, and $N_{\text{total\_biased}}$ is the total number of originally biased samples in the dataset.

Un+STA: The Un+STA metric is a weighted combination of Unbias and STA, which provides an overall evaluation of the model's debiasing effectiveness while preserving style accuracy, calculated as follows:

$$\text{Un} + \text{STA} = \alpha \times \text{STA} + (1 - \alpha) \times \text{Unbias} \tag{7}$$

where α represents the proportion of biased samples in the original dataset, and $(1 - \alpha)$ represents the proportion of unbiased samples in the original dataset.

Content Preservation (SIM): We train a sentence vector generator using an unsupervised training method from SIMCSE (Gao et al., 2021) on the original biased texts. We then measure semantic preservation by calculating the cosine similarity between sentence vectors before and after debiasing, given by:

$$\text{SIM} = \cos(\mathbf{v}_{\text{before}}, \mathbf{v}_{\text{after}}) \tag{8}$$

where $\mathbf{v}_{\text{before}}$ is the sentence vector before debiasing and $\mathbf{v}_{\text{after}}$ is the sentence vector after debiasing.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1078-1094, Jinan, China, August 11-14, 2025.