Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception

Anonymous ACL submission

Abstract

The pervasive spread of misinformation and disinformation in social media underscores the critical importance of detecting media bias. While robust Large Language Models (LLMs) have emerged as foundational tools for bias prediction, concerns about inherent biases within these models persist. In this work, we investigate the presence and nature of bias within LLMs and its consequential impact on media bias detection. Departing from conventional approaches that focus solely on bias detection in media content, we delve into biases within the LLM systems themselves. Through meticulous examination, we probe whether LLMs exhibit biases, particularly in political bias prediction and text continuation tasks. Additionally, we explore bias across diverse topics, aiming to uncover nuanced variations in bias expression within the LLM framework. Importantly, we propose debiasing strategies, including prompt engineering and model fine-tuning. Extensive analysis of bias tendencies across different LLMs sheds light on the broader landscape of bias propagation in language models. This study advances our understanding of LLM bias, offering critical insights into its implications for bias detection tasks and paving the way for more robust and equitable AI systems[1].

1 Introduction

Detecting media bias (Yu et al., 2008; Iyyer et al., 2014; Liu et al., 2022) was crucial due to the pervasive spread of misinformation and disinformation on social media platforms, profoundly shaping public perception and decision-making processes. Recently, researchers have increasingly turned to robust LLMs as foundational tools for media bias prediction (Lin et al., 2024; Liu et al., 2024). Compared to non-pretrained neural models or less powerful language models, LLMs offer enhanced capabilities, yet with an increased risk of bias introduction, given their superior performance and widespread use in media analysis and bias detection. Consequently, there is a growing need to examine bias within the bias detection process itself (Fang et al., 2023; Urman and Makhortykh, 2023; Esiobu et al., 2023).

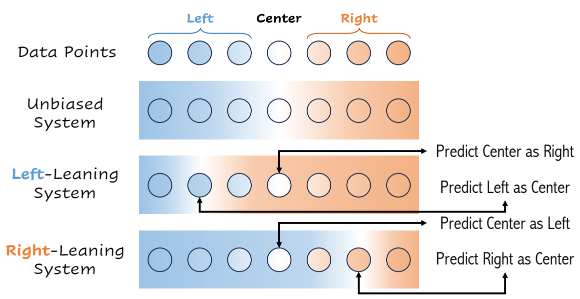


Figure 1: Interpretation of Biased Systems.

In this study, we investigate a series of research questions, including whether LLMs exhibit bias, their subsequent impact on media bias prediction results, a fine-grained analysis of LLM bias, and how debiasing affects performance. Before delving into our investigation, it's important to differentiate between the tasks of bias detection and LLM bias analysis. Bias detection in this context pertains to the media bias prediction task, which involves determining whether a given article exhibits bias. This task is text-oriented, focusing on analyzing input text. On the other hand, analyzing bias in LLM involves examining potential biases inherent in the LLM system itself, which is system-oriented and focusing on exploring biases within the system.

To better illustrate the impact of biased systems on media bias detection, Fig. 1 employs political bias prediction as an example. We observe that based on an unbiased system which is capable of accurately predicting the political ideology of given data points, the biased one may exhibit skewed predictions, leading to misinterpretations or misclassifications of the political ideology of the data points. In addition to this illustration, experiments reveal

[1]The code is available at `https://anonymous.4open.science/r/code-44B8`

that vanilla GPT-3.5 demonstrates an F1 score of 26.2% on FlipBias dataset (Chen et al., 2018) (a representative political bias prediction dataset), indicating its limited effectiveness in identifying the political leaning of articles. This raises the question of whether the unsatisfactory performance of LLMs in political ideology prediction stems from suboptimal capabilities inherent to LLMs or from inherent biases within the LLMs themselves.

We first explore the research question of whether LLMs exhibit political bias (RQ1) from two distinct perspectives: analyzing LLM bias through political bias prediction and text continuation tasks. The bias prediction perspective enables us to evaluate potential biases in an LLM's comprehension and prediction of specific given articles, while the text continuation perspective offers insights into the political leaning of LLMs' generated content when provided with a short prefix with pre-set political leaning. This yields broader implications of bias in LLMs for content generation applications.

Furthermore, unlike previous studies (Liu et al., 2021; Wambsganss et al., 2023a) that examines bias based on predefined dimensions such as demographics, gender, and location, we aim to explore the bias of LLMs at more granular and flexible levels. This involves examining bias at both predefined and latent topics to address the second research question: RQ2 Do LLMs exhibit consistent bias across topics? Further case examination of LLM bias under specific topics and proposed bias evaluation metrics reveal how biases vary across different topics. Through assessing bias consistency across topics that may vary temporally, we gain insights into how LLMs propagate biases.

Furthermore, we explore various debiasing methods, including isolating inherent bias through prompt engineering and adjusting the model's leaning via fine-tuning, to address the question: RQ3 How to debias LLMs and further improve performance? Throughout these investigations, we make several key observations that hold significance for future developments in LLM-based frameworks.

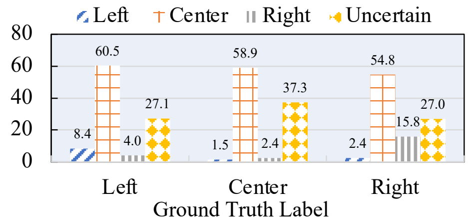
Lastly, we assess bias across different LLMs, both open-source and closed-source, to address the fourth research question: RQ4 Do various LLMs demonstrate similar bias tendencies? The results suggest that while different LLMs may demonstrate varying bias leanings, bias does indeed exist in the tested LLMs. Moreover, the performance of LLMs does not appear to correlate with the degree of bias exhibited by the models.

In summary, we provide a comprehensive investigation into the presence and nature of bias within LLMs and its consequential impact on media bias detection. The exploration of disparities between LLMs and human perception (i.e., the bias ground truth used in this work is labeled by humans) advances our understanding of LLM bias, offering critical insights into its implications for bias detection tasks and paving the way for more robust and equitable AI systems.

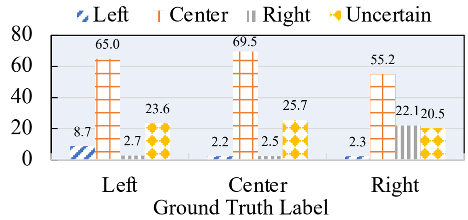
2 Related Work

Bias of LMs. Understanding bias within LMs is complex due to its normative and subjective nature, often influenced by various contextual and cultural factors (Gallegos et al., 2023). While providing a formal definition of bias can be challenging, it is commonly observed and studied through its manifestations in LM outputs. Biases manifest in various forms, including representational biases depicting certain social groups negatively (Beukeboom and Burgers, 2019), disparate system performance leading to misclassifications (Blodgett et al., 2016), and reinforcement of normativity (Bender et al., 2021). Misrepresentation of social groups can also exacerbate biases (Smith et al., 2022). While research (Hada et al., 2023; Gonçalves and Strubell, 2023; Conti and Wisniewski, 2023; Wang et al., 2023) has addressed bias in LMs broadly, our work focuses on political standing bias, aiming to elucidate discrepancies between LM cognition and human perceptions.

Bias Mitigation. Bias mitigation techniques encompass pre-processing, in-training, intra-processing, and post-processing interventions (Gallegos et al., 2023). Pre-processing involves altering model inputs, such as data and prompts (Venkit et al., 2023), to create more representative training datasets through techniques like data augmentation (Qian et al., 2022), data filtering (Garimella et al., 2022), prompt modification (Venkit et al., 2023), and debiasing pre-trained representations. Intra-processing methods (Zayed et al., 2023) modify model behavior at inference without further training, including decoding strategies, post hoc model adjustments, and modular debiasing networks. In-training techniques aim to reduce bias by modifying the optimization process, such as adjusting loss functions (Liu et al., 2021), updating probabilities, freezing parameters (Gira et al., 2022), or neuron removal (Joniak and Aizawa, 2022) during training.



(a) Political Bias Prediction on FlipBias



(b) Political Bias Prediction on ABP

Figure 3: LLM's prediction on FlipBias and ABP.

Post-processing (Tokpo and Calders, 2022) mitigates bias in model outputs through techniques like identifying and replacing biased tokens without altering original model parameters.

3 RQ1: Do LLMs exhibit political bias?

Previous work Rozado (2023) conducted 15 different political orientation tests on ChatGPT. The findings reported by Rozado (2023) reveal that ChatGPT tends to exhibit a preference for left-leaning viewpoints in its responses to questions. However, it is noteworthy that their investigations were based on a limited number of political orientation tests (i.e., 15 tests). In this section, we employ various bias analysis methods to further investigate the political bias exhibited by LLMs.

3.1 LLM-based Bias Prediction

We adopt vanilla ChatGPT model to conduct political leaning prediction on two popular datasets (i.e., FlipBias (Chen et al., 2018) and ABP (Baly et al., 2020)). The statistic of these two datasets can be found in Table 1. We can see that there are 1022 triples (i.e., each triple is with left-, center-, right-leaning article on same event) in FlipBias and more than 30k instances in ABP dataset. For each instance, we prompt `gpt-3.5-turbo-0613` with the following instruction to get the bias prediction results of vanilla ChatGPT:

> *Given the article provided below:* *TEXT ARTICLE*
> *Analyze the text content and assign a label from {left, right, center, uncertain}. In this context, "left" indicates a left-leaning article, "right" signifies a right-leaning article, "center" implies no obvious political leaning, and "uncertain" denotes that the political orientation could not be determined. Please provide your analysis and output a new single line containing only the assigned label.*

We present the bias prediction results in Fig. 3, comparing the ground truth labels (left, center, right) with the model's predictions (left, center, right, uncertain). Before delving into the analysis of the results in Fig. 3, we establish the following assumption. $\mathcal{A}0$: LLMs exhibit inherent political cognitive bias rather than an overall inability to judge articles' political leaning. $\mathcal{A}0$ implies that the prediction results of LLMs follow a linear bias pattern, as illustrated in Fig. 1. Based on the results in Fig. 3, we have the following observations:

- $\mathcal{O}1$: *The tested LLM exhibits left-leaning viewpoints.* By focusing on the proportions of Left-Center (where Left is the ground truth label and Center is the predicted label, e.g., the Left-Center proportion in Fig. 3(a) is 60.5) and Right-Center presented in Fig. 3, we observe that the Left-Center values surpass the Right-Center values on both datasets. These higher values indicate that the tested LLM demonstrates a left-leaning political cognitive bias, resulting in a higher likelihood of predicting left-leaning articles as centered articles. Furthermore, by comparing the Center-Left and Center-Right values across two datasets, we observe that the tested LLM tends to predict the centered article slightly more as right-leaning rather than left-leaning. This observation is consistent with the notion that the tested LLM exhibits left-leaning viewpoints.
- $\mathcal{O}2$: *Despite left-leaning tendencies, the tested LLM excels in predicting right-grounded articles.* An examination of the proportions of Left-Left and Right-Right predictions in Fig. 3 reveals that the Right-Right proportions are significantly higher than those of Left-Left. This suggests that the tested LLM excels in accurately classifying articles with a right-leaning perspective.

By comparing the results predicted by LLMs, we derive initial observation $\mathcal{O}1$, which is consistent with the findings reported by Rozado (2023). In the following, we explore the viewpoint leaning of LLMs through Article Continuation experiments and two distinct analytical approaches.

Dataset	Bias Label			# of Instances	Avg Length	Source
FlipBias (Chen et al., 2018)	Left 33.3%	Center 33.3%	Right 33.3%	3,066	1,077	New York Times, Huffington Post, Fox News and Townhall
ABP (Baly et al., 2020)	Left 34.5%	Center 36.6%	Right 28.8%	37,554	1,095	

Table 1: Statistics of FlipBias and ABP.

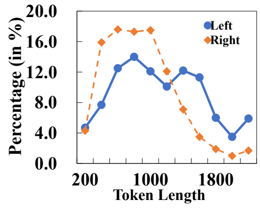


Figure 2: FlipBias Len.

3.2 LLM-based Article Continuation

Beyond the prediction-based analysis outlined earlier, we investigate LLM bias through article continuation. By supplying LLMs with prefixes derived from political articles and prompting them to extend these prefixes, we assess the political leaning of the generated suffix to analyze the inherent bias of LLMs. Our evaluation employs two methods for determining the political leaning of generated content: intuitive embedding-based similarity matching and Left and Right Vocabulary-based matching, following the approach proposed by (Fang et al., 2023; Wambsganss et al., 2023b).

Following this, we begin by providing a detailed description of the continuation implementation and then proceed to conduct in-depth examinations of bias in LLMs based on two distinct methods for determining political leaning of continued content.

Article Continuation. We prompt `gpt-3.5-turbo-0613` with a continuation prompt to generate text based on the given prefix.

> *Continue the text provided below:*
> *TEXT ARTICLE*

Building on the core idea of assessing the generated suffix to reflect the leaning of LLMs, we explore two automated methods to determine the bias label of the generated content.

Embedding-Based Similarity Matching. We utilize an off-the-shelf text embedding API of ChatGPT to create a vector database following (Peng et al., 2023). Specifically, the vector database comprises embeddings of all instances from the FlipBias dataset. For each instance in the FlipBias dataset, we construct prefixes (e.g., prefixes with a fixed number of tokens such as 20, 40, etc.) and obtain the continued suffix by prompting ChatGPT with the previously introduced prompt. Subsequently, we label the continuation suffix by calculating the similarity between the generated suffix and tripled instances[2] (i.e., left-leaning, center-leaning, and right-leaning articles) centered around the same event. We determine the bias label of the generated text based on the label of the instance with the highest similarity score. The entire process is formally described as follows.

[2]The triples are adjusted to match the length of the prefix, considering a prefix of length n, resulting in a length minus n.

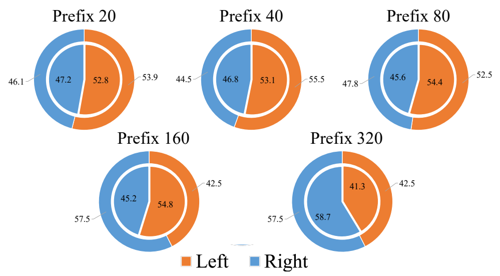


Figure 4: Article Continuation Results on FlipBias: The inner pie chart presents the outcomes of embedding-based similarity matching, while the outer doughnut illustrates the results of vocabulary-based matching.

$$\text{Similarity}_i = \frac{v_{\text{suffix}} \cdot v_i}{|v_{\text{suffix}}||v_i|}, \quad i \in \{\text{left, center, right}\} \tag{1}$$

$$\text{Bias Label} = \text{argmax}(\text{Similarity}_i) \tag{2}$$

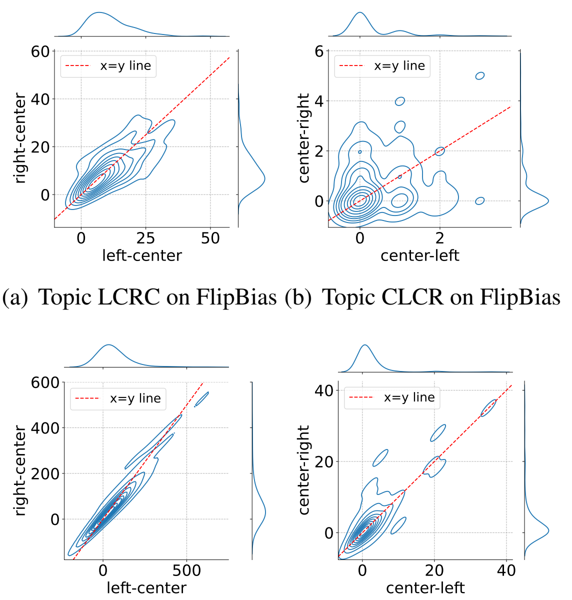
where $v_{(.)}$ represents the embedding of the text.

Left and Right Vocabulary-Based Matching. By following Yano et al. (2010), we first construct two vocabularies for left-and right-leaning articles separately. Each vocabulary is constructed by doing statistic of the word frequency for articles with ground-truth left and right labels and removing stop words (details are shown in Appendix A), which can represent the characteristic of the respective political leaning. The presence of a higher number of words from a specific vocabulary within an article indicates the alignment of the article with the corresponding political leaning. For instance, an article featuring more tokens from the left-leaning vocabulary indicates its left-leaning orientation.

In Fig. 4, we present the outcomes of article continuation experiments with varying prefix lengths (e.g., 20, 40 tokens) employing both embedding-based and vocabulary-based matching. It's important to note that only the relative percentages of left and right are presented, disregarding the center

Dataset	# of Topics	Latent	Avg Instance # Per Topic
FlipBias	152	✔	82
ABP	108	✗	348

Table 2: Statistics of Topics in FlipBias and ABP.



(a) Topic LCRC on FlipBias (b) Topic CLCR on FlipBias

(c) Topic LCRC on ABP (d) Topic CLCR on ABP

Figure 5: Joint plot displaying kernel density estimates.

situation. From Fig. 4, we can see that across prefix lengths ranging from 20 to 80, both label matching methods consistently show a higher percentage of left predictions, suggesting a left-leaning trend in continued articles. However, as the prefix length increases to 320, both methods begin to predict continued articles as more right-leaning. This change may be attributed to the fact that the average length of right-leaning articles is shorter than left-leaning articles (refer to Fig. 2). Therefore, when given a prefix with 320 tokens, the political leaning of the prefix becomes clearer, representing a substantial portion—approximately 40%—of the average length of Right articles (794 tokens) and 28% of Left articles (1111 tokens). This clearer representation of political leaning in the prefix makes it more likely for the LLM to generate a right-leaning suffix. Consequently, LLMs may find it easier to predict right-leaning continued suffixes.

4 RQ2: Do LLMs demonstrate consistent bias across all topics?

As elaborated in §3.1, our tested LLM exhibits a left-leaning bias compared to viewpoints derived from the ground-truth labels assigned by human evaluators. In this section, we delve into whether the LLM consistently showcases a leaning across all discussed topics. While the ABP dataset includes topic information, the FlipBias dataset lacks such information inherently. To address this, we construct latent topics following the methodology proposed by (Lin et al., 2024). The detailed process of latent topic construction is provided in Appendix B. As the constructed latent topics of FlipBias dataset are not predefined, we attempt to demonstrate their relevance and coherence to predefined topics in ABP dataset. This is achieved by presenting statistics on the (latent) topics of both datasets in Table 2, and by plotting joint distributions of Left-Center (i.e., where the ground-truth label is left and the predicted label is center) and Right-Center, as well as Center-Left and Center-Right, accompanied by kernel density estimates in Fig. 5. It is evident that the joint plots of the FlipBias and ABP datasets exhibit similar patterns. The main difference arises in the distributions based on predefined topics (i.e., Fig. 5(c) and Fig. 5(d)), which appear more focused compared to the distributions based on latent topics (i.e., Fig. 5(a) and Fig. 5(b)), which demonstrate greater dispersion.

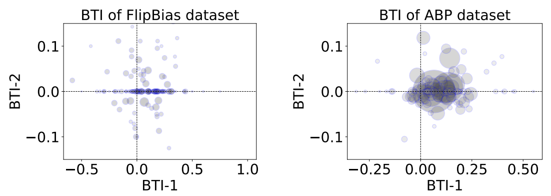
Visualization Based on Bias Tendency Index. Before presenting the results of viewpoints leaning in LLMs, we introduce two Bias Tendency Index (BTI) as follows.

$$\text{BTI-1} = \frac{\text{Count(left-center)}}{\text{Count(left)}} - \frac{\text{Count(right-center)}}{\text{Count(right)}} \quad (3)$$

$$\text{BTI-2} = \frac{\text{Count(center-right)}}{\text{Count(center)}} - \frac{\text{Count(center-left)}}{\text{Count(center)}} \quad (4)$$

where BTI-1 measures the bias tendency of the tested LLM regarding left and right-ground truth labeled articles. It quantifies the difference in predicting articles as center when the ground truth is left versus right. Similarly, BTI-2 focuses on the bias tendency of the LLM concerning articles with a ground truth label of center. It measures the disparity in predicting articles as right or left when the ground truth is center. A positive BTI-1 (BTI-2) suggests the tested LLM shows a left-leaning viewpoints, while a negative value suggests a right-leaning bias of LLM.

We present the distribution of BTI-1 and BTI-2 for the FlipBias and ABP datasets in Fig. 6. Each point in Fig. 6 represents a distinct topic, larger points indicate more instances located in the corresponding topic, and darker regions imply more topics located in that region. From Fig. 6, we find:



(a) Distribution on FlipBias (b) Distribution on ABP

Figure 6: BTI distribution of Topics on FlipBias and ABP. Darker colors and larger circles indicate more instances under the corresponding topic.

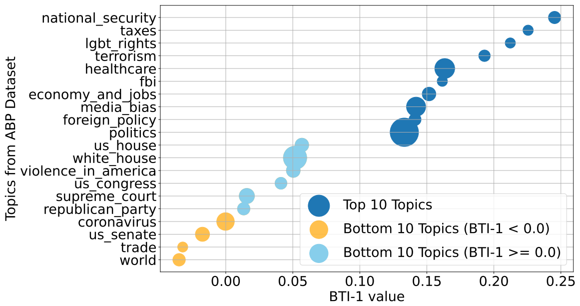


Figure 7: BTI-1 distribution for Top and Bottom 10 topics (ranked by BTI-1) with above-average frequency (i.e., the number of instances under respective topic).

• $\mathcal{O}3$: *The tested LLM does not exhibits same viewpoint leaning on all topics.* As discussed in §3 (i.e., $\mathcal{O}1$), the tested LLM demonstrates an overall left-leaning viewpoint on both the Flipbias and ABP datasets. By presenting the BTI-1 and BTI-2 values (where a positive value indicates left-leaning, referring to the explanation to Eq. (3) and Eq. (4)) for all topics in Figure 4, it is evident that while most points are situated in the right region of the figure (i.e., BTI-1 > 0), there are topics with notably negative values, indicating that the tested LLM displays right-leaning viewpoints on these topics.

• *The distribution of BTI-1 is more pronounced compared to the BTI-2 value*. Both Fig. 6(a) and Fig. 6(b) exhibit clear left-leaning tendencies in the distribution of BTI-1. While the distribution of BTI-2 on these two datasets appears more evenly spread, with points displaying both negative and positive BTI-2 values generally at similar scales.

• *The topic frequency distribution on FlipBias appears more evenly distributed compared to that of the ABP dataset.* By examining the sizes of points in Fig. 6(a) and Fig. 6(b), it is apparent that the clustered latent topics of FlipBias are more evenly distributed, indicating a balanced number of instances contained within each cluster. We provide interpretations of some clustered latent topics and the contained indicators in Appendix B.

Case Study of Biased Topics. To further analyze the LLM's leaning across various topics, we utilize several cases from the FlipBias and ABP datasets to demonstrate the relationship between viewpoint leaning and topic. For a more representative analysis, we select topics with above-average frequency and then rank them based on the calculated BTI-1 values. We present the top 5 and bottom 5 latent topics from FlipBias in Table 3. The interpretation of latent topics is obtained by prompting ChatGPT to provide a summary of the cluster indicators. More latent topic cases ranked by BTI-2 values can be found in Appendix B.

From Table 3, we observe that the trend of BTI-2 values is more centered around 0.0 when the range of BTI-2 extends to ± 0.5, which is consistent with the observation of Fig. 6. The LLM's left-leaning viewpoints on topics (upper part of Table 3) like journalism's use of citations, Obama's policies, and immigration critique reflect values of transparency, inclusivity, and social justice. This aligns with the narrative often seen in left-leaning media, emphasizing fact-checking, diverse perspectives, and human rights advocacy. These viewpoints may shaped by the model's training data and structural biases. The prevalence of Trump-related topics among the bottom 5 latent topics (lower part of Table 3) with negative BTI-1 suggests a potential right-leaning bias in the language model's treatment of Trump administration subjects. Given FlipBias's data collection primarily from 2013 to 2018, a period marked by heightened political polarization, this alignment hints at a correlation between temporal context and exhibited biases.

We further plot the BTI-1 distribution of the Top and Bottom 10 topics (ranked by BTI-1 values) with above-average frequency for the ABP datasets in Fig. 7. Upon closer examination, notable similarities emerge between topics with extreme values in both the Flipbias and ABP datasets. The analysis reveals similarities between extreme value topics in both Flipbias and ABP datasets, with positive values often focusing on security and terrorism, and negative values frequently discussing Trump's government and the US-China trade war. Given that ABP dataset's data is collected between 2019-2020, we infer that short-term hot topics like coronavirus tend to exhibit negative bias, while broader subjects like LGBT rights trend positively. The concentration of articles in the middle range of topics suggests that data scale may influence bias trends, with widely discussed topics reflecting human per-

Interpretation of Top and Bottom 5 latent topics (ranked by BTI-1 values)	BTI-1	BTI-2	Frequency
Comprehensive Use of Quotes and Citations in Journalism	0.44	0.00	81
Diverse Perspectives on President Obama's Policies and Actions	0.31	0.00	99
Analysis of Recent Terrorist Attacks and Security Measures in Various Cities	0.31	0.00	103
Critique of DACA Amnesty Program and Advocacy for Stricter Immigration Policies	0.29	0.00	89
Diverse Rhetorical Strategies in Political Discourse	0.29	0.00	97
Trump's Clashes with Federal Law Enforcement and Media	-0.17	0.11	80
Analysis of Media Coverage Surrounding Trump's Ratings, Criticisms, and Mental Fitness	-0.20	-0.03	98
Trump Administration's Response to Russia Sanctions and Political Fallout	-0.27	0.04	71
Satirical Commentary and Critique on Political Events and Figures	-0.30	0.00	58
Media Coverage of Trump Administration	-0.58	0.02	70

Table 3: Interpretation of Top and Bottom 5 Latent Topics on FlipBias.

spectives more closely.

5 RQ3: How to debias LLMs and further improve performance?

In this section, we delve into effective strategies for debiasing LLMs from two perspectives: strategies in in-context learning and fine-tuning methods for debiasing LLMs. While debiasing exploration is not the primary focus of this work, we present some results of basic debiasing attempts and assess debiasing effectiveness across three aspects: (1) overall debiasing results, (2) elimination of bias across all topics, and (3) potential performance improvements in bias prediction with debiased LLMs.

5.1 Prompts to Isolate Bias

Bias Label Explanation (BLE). As we mainly focus on political bias in this work, we incorporate detailed explanations of left-leaning and right-leaning perspectives sourced from Wikipedia to explore if adding bias label explanations to the prompt would help debiasing.

Few-shot Instruction. A small number (e.g., 3, 6, 12 in our experiment) of examples are explicitly designed to instruct LLMs to recognize biases. Given that FlipBias and ABP operate at the article level, we provide only the headlines to avoid lengthy inputs, as headlines have been proven effective in indicating bias (Chen et al., 2018). Further details of the chosen instances are in Appendix C.1.

Debiasing Statement (DS). Ganguli et al. (2023) discovered that providing a simple instruction to an LLM, urging it to avoid biases in its responses, can effectively mitigate inherent biases. Following Kaneko et al. (2024), we add a debiasing statement to our prompts: "*Please ensure that your answer is unbiased and free from reliance on stereotypes.*".

Models	BTI-1	BTI-2	Pre	Rec	BiF1	MiF1	MaF1
Vanilla	0.06	0.01	89.0	15.4	26.2	42.3	39.4
BLE	0.03	0.00	89.3	9.4	17.0	38.8	34.3
3-shot	0.06	0.00	93.1	11.3	20.2	40.3	36.3
6-shot	0.04	0.00	92.6	9.7	17.6	39.3	34.8
9-shot	0.04	0.00	**96.9**	7.7	14.3	38.3	33.1
DS	0.01	0.00	91.9	6.7	12.4	37.4	31.8
L-FT	0.00	-1.00	66.7	**100.0**	**80.0**	**66.7**	40.0
LC-FT	-0.17	-0.41	67.8	43.0	52.6	48.4	48.0
LCR-FT	-0.00	-0.68	68.6	89.9	77.8	65.8	**51.7**

Table 4: Debiasing results on FlipBias.

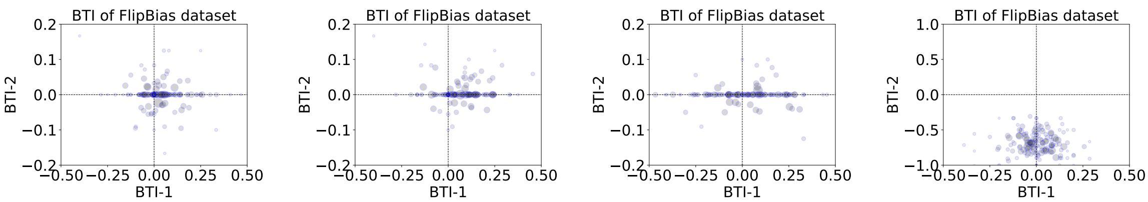
5.2 Fine-Tuning to Debias

By observing the results of Fig. 3, we infer that the LLM demonstrates better performance in clarifying right-label articles from center-label articles compared to clarifying left-label articles from center-label ones. This observation suggests a potential deficiency in the LLM's ability to accurately recognize left-leaning evaluation criteria. To address this, we adjust the proportion of left-leaning articles in the fine-tuning instances to investigate how varying proportions impact the debiasing process. Specifically, we fine-tune `gpt-3.5-turbo` using 300 labeled instances (sampled from the regular training sets of datasets) with three different proportions: all left-label articles (L-FT), a mixture of left-label and center-label articles (LC-FT), and an equal distribution of left-label, center-label, and right-label articles (LCR-FT).

5.3 Assessment of Debiasing Strategies

We evaluate the debiasing methods introduced in §5.1 and §5.2 in this subsection. Apart from BTI, the other metrics follow Lin et al. (2024).

General Leaning and Bias Prediction Performance Comparison. The debiasing results on FlipBias are reported in Table 4. We observe that while finetuning methods generally exhibit better bias prediction performance gains (e.g., better BiF1



(a) BLE (b) 3-Shot (c) DS (d) LCR-FT

Figure 8: BTI distribution of Topics on FlipBias dataset after debiasing. More distributions are in Appendix D.

Models	BTI-1	BTI-2	Pre	Rec	BiF1	MiF1	MaF1
LLaMa2	0.04	0.25	72.7	47.1	57.2	52.7	52.2
Vicuna	-0.01	0.07	68.0	19.1	29.8	39.8	38.5
Mistral	0.00	-0.57	69.9	84.2	76.4	65.3	55.4
GPT-3.5	0.06	0.01	89.0	15.4	26.2	42.3	39.4
GPT-4	0.06	-0.04	85.1	30.3	44.7	50.0	49.5

Table 5: Comparison results of different LLMs.

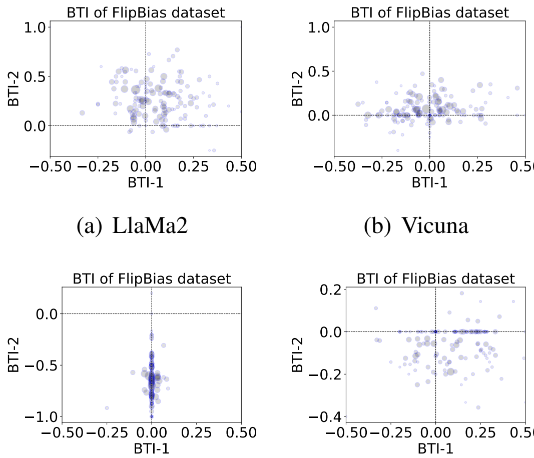
and MaF1), they also introduce more bias to the finetuned LLMs, as reflected by larger BTI-1 or BTI-2 values after finetuning. On the other hand, prompt-based debiasing methods show impressive effects, especially DS (Ganguli et al., 2023), which is extremely easy yet effective.

Topic-Level Bias Comparison. We further display the bias tendency index (BTI) distribution on FlipBias after applying some representative debiasing methods in Fig. 8, while distributions of additional debiasing methods and results from the ABP dataset can be found in Appendix D. From Fig. 8, we observe that prompt engineering-based debiasing shows better results, as reflected in the BTI values for topics being centered around 0.0, which is consistent with the general performance comparison results we introduced in the last paragraph. Additionally, the overall shift in the BTI distribution after LCR-FT debiasing, as shown in Fig. 8(d), indicates that finetuning LLMs may result in better performance (refer to bias prediction results reported in Table 4), but it may inadvertently introduce more severe bias.

6 RQ4: Do various LLMs exhibit similar bias tendencies?

In the previous sections, we conduct experiments using a representative LLM named GPT-3.5. In this section, we extend our analysis to include biases of additional LLMs, both closed-source and open-source. These include `Llama-2-7B-Chat`, `Vicuna-7B-v1.5`, `Mistral-7B-v0.1`, and `gpt-4-0125-preview`.

We present the bias prediction results and BTI values of these LLMs in Table 5, along with the topic-level BTI distribution in Fig. 9. From Table 5, it can be observed that LLaMa2 and Mistral even show better political bias performance than GPT-3.5 and GPT-4. However, it is important to clarify that although LLaMa2 and Mistral exhibit better performance according to current classification metrics, they display severe issues such as denying answering and generating unrelated content instead of predicting bias labels (for about 20% of the testing). Additionally, considering the bias index BTI-1 and BTI-2 values, almost all LLMs exhibit bias, with Mistral showing a general right-leaning tendency, which differs from other LLMs. The fine-grained bias distribution in Fig. 9 is consistent with the overall bias reported in Table 5.



(a) LlaMa2 (b) Vicuna (c) Mistral (d) GPT4

Figure 9: BTI of Topics on FlipBias for various LLMs.

7 Conclusion

In summary, our investigation reveals inherent biases within LLMs and their significant impact on media bias detection. Departing from conventional approaches, we explore biases within LLM systems themselves, particularly in political bias prediction task. Our findings highlight the need for debiasing strategies and provide insights into the broader landscape of bias propagation in language models.

Limitations

This work is subject to limitations in two main aspects: (1) Limited Focus on LLM Bias in Media Bias Prediction: The scope of bias analysis is constrained by the availability of three-way (left-, center-, and right-leaning) labeled data. Our study relies on two political bias prediction datasets with three-way labels to investigate biases during LLM prediction. However, datasets with only biased and non-biased labels would not suffice for our analysis in this paper. (2) Assumption of Ground Truth: We operate under the assumption that human-labeled data serves as an unbiased ground truth for assessing LLM biases. Nevertheless, human annotations are inherently subjective and may be influenced by individual biases, potentially impacting the validity of our evaluations.

References

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. *arXiv preprint arXiv:2010.05338*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Camiel J Beukeboom and Christian Burgers. 2019. How stereotypes are shared through language: a review and introduction of the aocial categories and stereotypes communication (scsc) framework. *Review of Communication Research*, 7:1–37.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*.

Wei-Fan Chen, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Learning to flip the bias of news headlines. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 79–88, Tilburg University, The Netherlands. Association for Computational Linguistics.

Lina Conti and Guillaume Wisniewski. 2023. Using artificial French data to understand the emergence of gender bias in transformer language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10362–10371, Singapore. Association for Computational Linguistics.

David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. 2023. Robbie: Robust bias evaluation of large generative language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3764–3814.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2023. Bias of ai-generated content: an examination of news produced by large language models. *arXiv preprint arXiv:2309.09825*.

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

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Aparna Garimella, Rada Mihalcea, and Akhash Amarnath. 2022. Demographic-aware language model fine-tuning as a bias mitigation technique. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 311–319.

Michael Gira, Ruisu Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69.

Gustavo Gonçalves and Emma Strubell. 2023. Understanding the effect of model compression on social bias in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2663–2675.

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 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1862–1876.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122.

Przemyslaw Joniak and Akiko Aizawa. 2022. Gender biases and where to find them: Exploring gender bias in pre-trained transformer-based language models using movement pruning. *arXiv preprint arXiv:2207.02463*.

Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. 2024. Evaluating gender bias in large language models via chain-of-thought prompting. *arXiv preprint arXiv:2401.15585*.

Luyang Lin, Lingzhi Wang, Jinsong Guo, Jing Li, and Kam-Fai Wong. Inditag: An online media bias analysis and annotation system using fine-grained bias indicators.

Luyang Lin, Lingzhi Wang, Xiaoyan Zhao, Jing Li, and Kam-Fai Wong. 2024. Indivec: An exploration of leveraging large language models for media bias detection with fine-grained bias indicators. *arXiv preprint arXiv:2402.00345*.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. 2024. Teller: A trustworthy framework for explainable, generalizable and controllable fake news detection. *arXiv preprint arXiv:2402.07776*.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nick Beauchamp, and Lu Wang. 2022. Politics: pretraining with same-story article comparison for ideology prediction and stance detection. *arXiv preprint arXiv:2205.00619*.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: Measuring chatgpt political bias. *Public Choice*, 198(1):3–23.

Ruoling Peng, Kang Liu, Po Yang, Zhipeng Yuan, and Shunbao Li. 2023. Embedding-based retrieval with llm for effective agriculture information extracting from unstructured data. *arXiv preprint arXiv:2308.03107*.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*.

David Rozado. 2023. The political biases of chatgpt. *Social Sciences*, 12(3):148.

David Rozado. 2024. The political preferences of llms. *arXiv preprint arXiv:2402.01789*.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in llm simulations of debates. *arXiv preprint arXiv:2402.04049*.

Ewoenam Kwaku Tokpo and Toon Calders. 2022. Text style transfer for bias mitigation using masked language modeling. *arXiv preprint arXiv:2201.08643*.

Aleksandra Urman and Mykola Makhortykh. 2023. The silence of the llms: Cross-lingual analysis of political bias and false information prevalence in chatgpt, google bard, and bing chat.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao'Kenneth' Huang, and Shomir Wilson. 2023. Nationality bias in text generation. *arXiv preprint arXiv:2302.02463*.

Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Neshaei, Roman Rietsche, and Tanja Käser. 2023a. Unraveling downstream gender bias from large language models: A study on AI educational writing assistance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10275–10288, Singapore. Association for Computational Linguistics.

Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Parsa Neshaei, Roman Rietsche, and Tanja Käser. 2023b. Unraveling downstream gender bias from large language models: A study on ai educational writing assistance. *arXiv preprint arXiv:2311.03311*.

Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023. A causal view of entity bias in (large) language models. *arXiv preprint arXiv:2305.14695*.

Tae Yano, Philip Resnik, and Noah A Smith. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 152–158.

Bei Yu, Stefan Kaufmann, and Daniel Diermeier. 2008. Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5(1):33–48.

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 AAAI Conference on Artificial Intelligence*, volume 37, pages 14593–14601.

A Left-Right Vocabulary Corpus Construction

We construct the Left-Right vocabulary corpus using the ABP dataset. Initially, all articles in ABP are tokenized using the NLTK Python package. Tokens are converted to lowercase and filtered using a stopwords corpus. Each token is then labeled based on the articles they appear in.

To create the Left-Right Vocabulary Corpus, we prioritize tokens labeled with significantly higher frequencies in either Left or Right articles. Specifically, we calculate the Left ratio by dividing a token's frequency in Left articles by the total tokens in Left articles, and similarly for the Right ratio. Tokens are included in the Left vocabulary list only if the Left ratio is more than twice the Right ratio.

From the Left vocabulary list, we select the top 2000 most frequent tokens. We then select 1295 tokens from the Right vocabulary list to match the total frequency sum of the Left tokens. This corpus is validated against the vocabulary of Yano et al. (2010). The constructed vocabularies will be publicly available for future research.

B Latent Topic Construction

Inspired by IndiVec (Lin et al., 2024), we prompted ChatGPT to construct fine-grained media bias indicators using the Flipbias dataset. These indicators summarize key points that may reflect media bias in each article. To organize the topics covered in these articles more effectively, we performed strict clustering through Hierarchical Clustering based on Euclidean distance applied to the indicators extracted from Flipbias. We utilized AgglomerativeClustering from the Scikit Learn package, setting the distance threshold to 2. The embeddings of the indicators were derived from OpenAIEmbeddings. Ultimately, 19,671 indicators were clustered into 152 clusters, each representing a latent topic.

Latent Topics and Corresponding Clustered Indicators Details of the clustered indicators are provided in Table 6.

Latent Topic Cases Ranked by BTI-2 Values Rankings of latent topic cases based on BTI-2 values are shown in Table 7.

C Implementation Details

C.1 Details of Prompts to Isolate Bias

In §5.1, we discussed the methods to debias LLMs. Here we provie details of these debaising methods.

Bias Label Explanation In Bias Label Explanation (BLE) method, we adopt explanations as listed in Table 8.

Few-shot Instances In Few-shot Instruction method, we randomly selected 4 Left-Center-Right triples from the dataset Flipbias and then used the titles as the instances of few-shot instruction, which are listed in Table 9.

C.2 More Finetuning Implementation Details

GPT Finetuning Details We fine-tuned gpt-3.5-turbo through the API supplied by OpenAI. 300 instances are randomly selected from the dataset ABP according to our setting as the training set. The hyperparameters of the number of epochs is 3 and the batch size is 32.

D Debiasing Results

We list the BTI distribution of Topics on ABP and FlipBias datasets after prompt debiasing in Fig. 10 and Fig. 11, separately.

E More Discussions

E.1 Distinguish from Related Works

Existing research has explored political bias in LLMs. Here, we differentiate our contributions from key related works:

(Taubenfeld et al., 2024) examines LLMs in simulating political debates, revealing conformity to inherent social biases despite specific directions. (Taubenfeld et al., 2024)focuses on interaction simulation, whereas our research centers on bias detection, emphasizing end-to-end and fine-grained analyses. (Rozado, 2024) analyzes bias through 11 political orientation tests, while our study highlights limitations of orientation tests and provides robust quantitative analysis based on extensive datasets, offering a broader perspective on LLM biases. (Urman and Makhortykh, 2023) investigates LLM-Chat Models' responses to predefined queries, focusing on non-responses and false responses. While related to the political domain, it primarily addresses jailbreaking and harmful effects. Our research questions are more specific, targeting systematic bias detection. (Motoki

Indicators	Interpreted Topic
"Provides figures and quotes from individuals involved in the issue", "The article cites statements from various individuals involved in the case, including lawyers, politicians, and advocacy groups.", "The text quotes various experts and government officials to support its claims.", "cites tweets and quotes from Trump, experts, and state officials to support the claims made", "Quotes from various food experts and diplomats.", "The article cites multiple sources, including government documents and quotes from officials."	Comprehensive Use of Quotes and Citations in Journalism
"Describes President Obama's decision as "benighted" and "cowardly" while praising President Trump's decision", ""swipes at Joe Biden," "knocks primary rival Bernie Sanders," "gripes about former President Barack Obama"", "frames the issue as a result of understaffing and mismanagement, blames the Obama administration, and highlights the need to protect the president", "Celebratory tone towards Obama, sarcastic and mocking tone towards Democrats", "Portrays Democrats as wanting a grander celebration, mocks Obama and the holiday", "Frames the decision as a potential unwinding of an Obama executive action, includes criticism from Democrats and environmental groups", "Describes the tough-on-crime approach as a reversal of Obama's "Smart on Crime" policy, implying a negative change"	Diverse Perspectives on President Obama's Policies and Actions
"Mentions celebrations and security measures in various cities", "Mentions specific incidents of terrorism and security measures in different cities", "Presents the incident as a terrorist attack and highlights the victims' nationalities", "Mentions previous vehicle attacks and quotes from witnesses", "Provides examples of other major music event attacks", "Mentions the arsenal of weapons and ammunition recovered, suggesting the possibility of an accomplice", "Mentions the London subway station fire as a terrorist incident caused by an improvised explosive device", "The article provides examples of previous attacks and the use of improvised explosive devices."	Analysis of Recent Terrorist Attacks and Security Measures in Various Cities

Table 6: Clustered Indicators and Interpreted Topics.

Interpretation of Top and Bottom 5 latent topics (ranked by BTI-2 values)	BTI-1	BTI-2	Frequency
Trump's Clashes with Federal Law Enforcement and Media	-0.17	0.11	80
Examining Controversial Tactics: Dissecting Allegations and Defenses in Recent Political Affairs	0.18	0.10	64
Analysis of Congressional Dynamics: Trump's Strategy, Witness Battles, and Financial Focus	-0.05	0.09	72
Bipartisan Cooperation in Senate: Struggles and Progress	-0.05	0.08	68
Unveiling the Constitutional Crisis: Examining Government Overreach and the Erosion of Rights	0.25	0.07	72
Understanding Textual Analysis: The Importance of Examples and Analogies	0.18	-0.05	59
Statewide Controversies: Voter Rights, Criminal Justice, and Transition Integrity	0.18	-0.06	60
Analyzing Political Discourse: Insights from Trump Administration and Beyond	-0.02	-0.08	76
Examining Biased Reporting in Political Discourse: Imbalance in State of the Union Addresses	0.20	-0.09	96
Critical Discourse Analysis of Media Portrayal on Trump's Governance	-0.01	-0.10	91

Table 7: Interpretation of Top and Bottom 5 Latent Topics on FlipBias.

et al., 2024) evaluates ChatGPT's responses to ideological questions. It focuses solely on ChatGPT, whereas our work encompasses a broader range of LLMs and addresses comprehensive research questions, providing a more extensive analysis.

These studies contribute to understanding political bias in LLMs. However, our work stands out by offering a more systematic exploration, addressing four comprehensive research questions, employing intricate experimental designs, and analyzing a broader range of LLMs, thus significantly extending the current body of research.

E.2 Exploration of Different Embeddings

In Section 3.2, we explored the embedding-based similarity matching method using embeddings from the GPT-3.5 model. Here, we extend our investigation to include another embedding source: sentence-t5-base[3] (T5-Base). The continuation results using T5-Base embeddings are summarized in Table 11. The calculation of left and right percentages in the table follows the methodology detailed in Figure 4.

From Table 11, we observe similar trends across different prefix lengths as shown in Fig. 4, although there are slight variations in predictions for prefix length = 320. Overall, the findings indicate a predominant left-leaning trend in continued articles, consistent with our earlier observations using GPT-3.5 embeddings.

[3] https://huggingface.co/sentence-transformers/sentence-t5-base

Left-wing politics describes the range of political ideologies that support and seek to achieve social equality and egalitarianism, often in opposition to social hierarchy as a whole or certain social hierarchies. Left-wing politics typically involve a concern for those in society whom its adherents perceive as disadvantaged relative to others as well as a belief that there are unjustified inequalities that need to be reduced or abolished through radical means that change the nature of the society they are implemented in.
Right-wing politics is the range of political ideologies that view certain social orders and hierarchies as inevitable, natural, normal, or desirable, typically supporting this position based on natural law, economics, authority, property or tradition. Hierarchy and inequality may be seen as natural results of traditional social differences or competition in market economies.
Centrism is a political outlook or position involving acceptance or support of a balance of social equality and a degree of social hierarchy while opposing political changes that would result in a significant shift of society strongly to the left or the right.

Table 8: Examples of Article Continuation

Text	Label
Trump Accuses His Justice Department, FBI Of Favoring Democrats	Left
Explosive memo released as Trump escalates fight over Russia probe	Center
Trump accuses FBI, DOJ leadership of bias against Republicans and in favor of Dems	Right
Shutdown truce just delays Trump's big dilemma	Left
Winners and losers from the government shutdown	Center
Centrists break Senate logjam, pave new path for 'common sense' bipartisanship	Right
North Korean insults to U.S. leaders are nothing new — but Trump's deeply personal reactions are	Left
Trump trades 'short and fat' barb with N Korea's Kim	Center
Trump Take To Social Media To Hit Back At 'Short and Fat' Kim Jong-un	Right
After 16 Futile Years, Congress Will Try Again to Legalize 'Dreamers'	Left
The clock is ticking': Graham and Durbin urge action on bipartisan DREAM Act by the end of September	Center
Republican Sen. Cory Gardner agrees to support bipartisan Dream Act after Trump rescinds DACA	Right

Table 9: few-shot instances

E.3 Article Continuation Examples

In §3.2, we adopt GPT-3.5 to conduct article continuation. We first report the average suffix length for each setting as follows: 490.1, 487.5, 479.0, 463.7, and 473.9 for prefixes with lengths of 20, 40, . . . , 320, respectively. Due to the strong capability of GPT-3.5, the generated suffixes are quite consistent with the prefix. Table 10 shows the randomly selected examples in different prefix settings of article continuations.

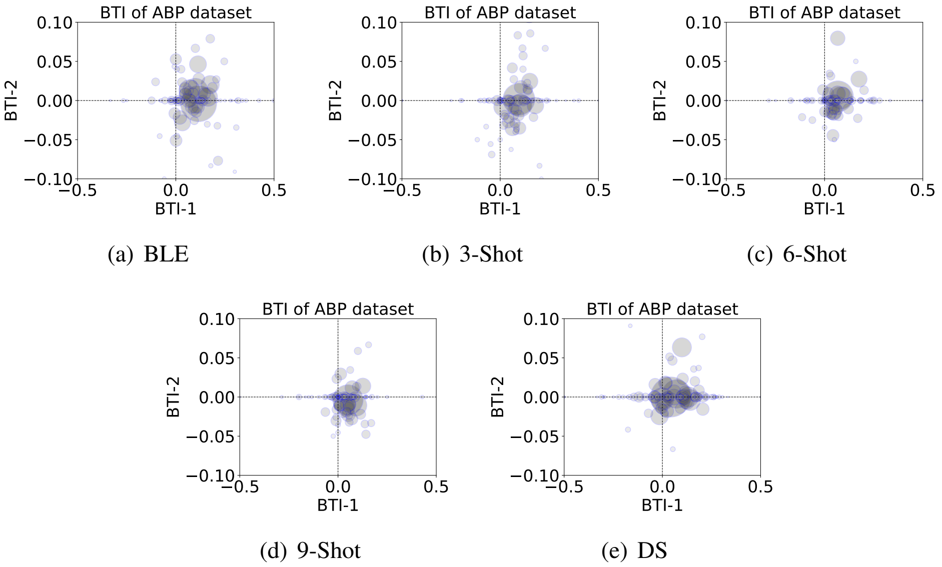
E.4 Finetuning Other LLMs

In addition to the finetuning debiasing results of `ChatGPT 3.5` reported in Table 4, we examined the finetuning debiasing method on a smaller LLM, specifically `LLaMa-2-7B-Chat`.

We report the BTI-1 and BTI-2 scores for LLaMa2 in Table 12, where:

- **LLaMa2:** `LLaMa-2-7B-Chat` without finetuning.
- **LLaMa2 LCR-FT:** `LLaMa-2-7B-Chat` finetuned according to the setting described in Section 5.2, with 300 articles evenly distributed among left-label, center-label, and right-label categories (LCR-FT).
- **LLaMa2 Finetune (Right Leaning data):** `LLaMa-2-7B-Chat` finetuned with 300 right-leaning data, where grounded center articles are labeled as left, and grounded right articles are labeled as center.

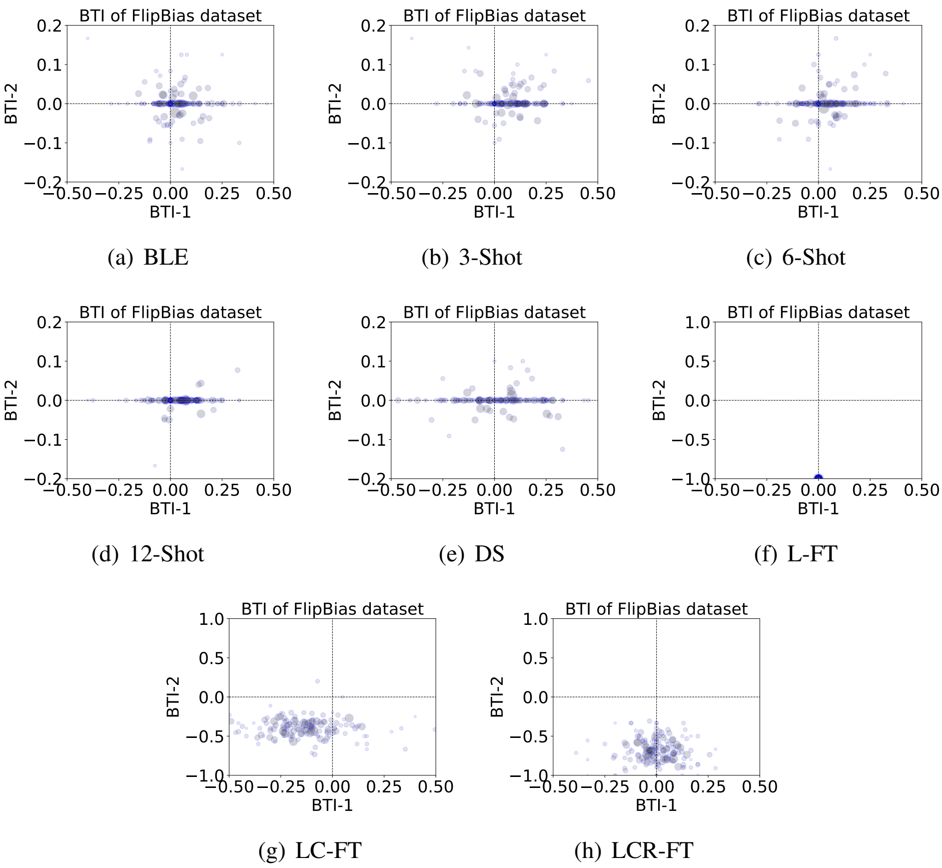
We further report the BTI distribution of `LLaMa2` and `LLaMa2 LCR-FT` in Fig. 12. We observe that although the averaged BTI-1 scores do not exhibit significant changes in Table 12 before and after finetuning, upon examining the topic-level distribution (refer to Fig. 12), we notice a more centralized BTI-1 distribution.



(a) BLE (b) 3-Shot (c) 6-Shot

(d) 9-Shot (e) DS

Figure 10: BTI distribution of Topics on ABP dataset after prompt debiasing.



(a) BLE (b) 3-Shot (c) 6-Shot

(d) 12-Shot (e) DS (f) L-FT

(g) LC-FT (h) LCR-FT

Figure 11: BTI distribution of Topics on FlipBias dataset after debiasing.

Prefix-20 Example:
Prefix: WASHINGTON (Reuters) - President Donald Trump and his Republican allies in Congress on Friday escalated a campaign against
Suffix: Special Counsel Robert Mueller and the ongoing Russia probe, as an explosive memo was released that alleges misconduct within the FBI and Justice Department. The memo, written by Republican lawmakers, accuses the FBI of abusing its surveillance powers in obtaining a warrant to spy on a former Trump campaign aide…

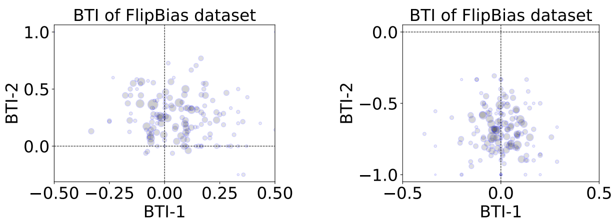
Prefix-40 Example:
Prefix: WASHINGTON (Reuters) - President Donald Trump and his Republican allies in Congress on Friday escalated a campaign against U.S. law enforcement agencies over their probe of Trump's ties to Russia, releasing a disputed memo that
Suffix: accuses the FBI of bias and misconduct. The explosive memo, which was released by the House Intelligence Committee, has sparked intense controversy and further deepened the divide between the Trump administration and the intelligence community. …

Prefix-80 Example:
Prefix: … Trump's ties to Russia, releasing a disputed memo that the FBI warned was misleading and inaccurate. The previously classified document, written by Republicans on the House of Representatives intelligence committee, alleges that the federal probe of potential collusion between Trump's 2016 presidential campaign and
Suffix: Russian officials was politically motivated and based on questionable evidence. The release of the memo marks a significant turning point in the ongoing battle between Trump and the intelligence community. Trump has long criticized the investigation into Russian interference in the 2016 election, calling it a "witch hunt" and claiming that it is a politically motivated attempt to undermine his presidency. …

Table 10: Examples of Article Continuation

	20 (left/right)	40 (left/right)	80 (left/right)	160 (left/right)	320 (left/right)
GPT3.5	52.8/47.2	53.1/46.8	54.5/45.6	54.8/45.2	41.3/58.7
T5-base	50.0/50.0	49.9/50.1	52.8/47.2	51.7/48.3	57.6/42.4

Table 11: Comparison of Two Embeddings (GPT3.5 v.s. T5-Base) Results in Article Continuation Experiments



(a) LlaMa2 (b) LlaMa2 LCR-FT

Figure 12: BTI distribution of Topics on FlipBias dataset for `LlaMa2` and `LlaMa2 LCR-FT`.

Model	BTI-1
LLaMa2	0.04
LLaMa2 LCR-FT	-0.024
LLaMa2 Finetune (Right Leaning data)	0.02

Table 12: BTI-1 of LLaMa2 and Finetuning Methods