Evaluating and improving the performance of Debiasing using Causal-Guided Active Learning for Large Language Models

Akshita Gupta[1], Krishnaprasad Vijayshankar[1], Mahita Kandala[1]

[1]Carnegie Mellon University,

Abstract

The use of LLMs as a source of knowledge and problem solving has increased significantly in the past year. This growing use also brings with it problems such as bias in LLMs. Research has shown that this bias extends to more than just social biases present in the datasets. Even complex models are prone to intrinsic biases such as positional bias in MCQs and verbosity bias in more reasoning based questions. (DU et al., 2024) aim to use Active Learning to identify the most bias inducing data points and further, cluster bias patterns and design targeted prompts to counteract them. This gives promising results on LLaMA-2 13B, however with LLaMA-3 8B, an increase in position bias is observed. This paper aims to mitigate bias in more complex models in datasets using LLMs as a judge (Zheng et al., 2023) while also introducing alternative metrics like Weighted Circular Evaluation and Mean Swapped Accuracy to assess how well a model is debiased. The debiased model shows a 4% improvement in swapped accuracy and a 10% improvement in weighted circular evaluation in zero-shot prompting, and a 2.5% improvement in swapped accuracy and a 3% improvement in circular evaluation in few-shot prompting. These results demonstrate improvements over the baseline and highlight the effectiveness of CAL for more complex models.

1 Introduction

The growing use of Large Language Models (LLMs) has raised several concerns about their effectiveness and the ethical implications surrounding their outputs. Despite their ability to generate human-like text, LLMs often reflect biases present in their training data, leading to outputs that may reinforce harmful stereotypes or create unfair outcomes. Additionally, the effectiveness of LLMs in maintaining context, accuracy, and cultural sensitivity can vary significantly, presenting challenges in ensuring reliable and unbiased language generation across diverse applications.

These concerns have led to a surge in the research of detecting bias in LLMs and mitigating them through various strategies, while retaining the capabilities of LLMs to their full extent. (DU et al., 2024) explore the use of Active Learning to improve the debiasing process in LLMs. Causal Guided Active Learning (CAL) addresses the limitations of traditional debiasing strategies by integrating active learning with causal mechanisms to identify and understand bias patterns in LLMs. Active learning allows for the selection of informative training samples, while causal mechanisms help in detecting biases by evaluating the stability of model predictions across semantically similar inputs. This framework also leverages in-context learning (ICL) to further reduce the need for extensive fine-tuning, making it a cost-effective debiasing method.

This study focuses on improving CAL's performance by conducting a more detailed analysis of positional bias in the MT-Bench and Chatbot datasets. These datasets are particularly relevant for evaluating generative models due to their emphasis on ranking tasks and conversational contexts, where positional bias can significantly influence outputs. By focusing on positional bias in advanced models like LLaMA-3, this paper aims to improve the effectiveness of CAL as a debiasing framework while preserving the functional capabilities of LLMs. Additionally, it introduces novel evaluation metrics such as Weighted Circular Evaluation, and Mean Swapped Accuracy designed to more accurately capture the extent of bias mitigation, contributing to the development of fairer and more robust LLMs for diverse real-world applications.

This paper is organized into seven sections. Section 2 reviews prior work on identifying and mitigating bias in LLMs, with a particular emphasis on CAL as a debiasing framework. Section 3 presents

a preliminary analysis of CAL's performance on the MT-Bench dataset. Section 4 provides an overview of the methodology employed for measuring and mitigating bias. Finally, Sections 5 and 6 outline directions for future work and conclusion of the study respectively.

2 Related Works

Related literature works have made multiple efforts to develop techniques for efficient detection of bias in LLMs. Addressing bias first requires understanding of its nature and prevalence in LLMs. Amongst several types of biases that are prelevant, common types include gender biases, age biases, racial biases, sexual biases, and appearance-based biases (Navigli et al., 2023). These biases can emerge based on context, and while some LLMs may appear resistant to certain biases, they often surface more clearly when multiple social identities, such as race and gender, intersect. This intersectional bias highlights the complexities of bias in LLMs and its broader implications (Ungless et al., 2022).

Besides all the listed biases, LLMs also exhibit a distinct type of bias in question-answering (QA) scenarios, known as positional bias, where the model prioritizes content based on its position within the input context (Wang et al., 2024). A severe case of such bias is the anchored bias observed in the GPT-2 family, where the model consistently prefers certain answers, such as "A" (Li and Gao, 2024). This bias can significantly undermine the reliability of QA systems. There have been numerous attempts to mitigate Positional bias by data augmentation (He et al., 2023; Zhu et al., 2024), to ensemble-based debiasing methods (Clark et al., 2019; He et al., 2019).

Measuring positional bias cannot be simply achieved through accuracy, as accuracy alone is not a reliable metric for assessing a model's freedom from position-related bias. Several studies have proposed different metrics to evaluate how LLMs handle positional bias. (Zhu et al., 2024) introduce **positional consistency** which measures how often a model selects the same response before and after the order of options is reversed. While (Zheng et al., 2024) suggest measuring positional bias by analyzing the balance of recalls across different option IDs, using the **standard deviation of recalls (RStd)** as a quantitative metric. (Chhikara et al., 2024; Plecko and Bareinboim, 2023) also discuss the concept of **statistical parity** as a method for fairness evaluation in LLMs. Despite all these advancements, there is still no standardized benchmark for reliably evaluating positional bias in LLMs.

Recent research on mitigating biases in LLMs has increasingly leveraged causal inference techniques. Causal inference based methods effectively capture causal relationships in chain-of-thought reasoning, aligning with LLMs' representation space. (Zhang et al., 2024; Zhou et al., 2023). Causality-based methodologies not only improve accuracies but also offer a promising approach for mitigating biases in language models (Liu et al., 2024). (Zhou et al., 2023) apply interventions on non-causal factors in different demographic groups and introduce an invariant risk minimization loss to mitigate bias while preserving task performance. Similarly, (Wang et al., 2023) propose an enhancement algorithm using causal inference to improve the accuracy of biased models, particularly in imbalanced datasets. Incorporating active learning with causal inference methods further enhances this framework by autonomously identifying bias patterns, reducing manual effort, and improving scalability. Active learning also prioritizes informative examples, detects intersectional biases, and supports continuous adaptation, ensuring robust, fair NLP applications (DU et al., 2024).

2.1 Causal Guided Active Learning (CAL)

This section focuses on the approach proposed by (DU et al., 2024) highlighted in Figure 1. CAL uses causal invariance to identify bias patterns, and checks whether the model's predictions remain consistent for semantically similar sentences. When outputs differ for sentences with similar semantics, it indicates that the model has failed to capture invariant predictive information, leading to the identification of counterexample pairs. Bias detection is based on two main criteria: (i) the influential criterion, which focuses on low predicted probabilities and high discrepancy from gold outputs, and (ii) the typical criterion, which detects consistent bias patterns across counterexample pairs. This automated process eliminates the need for predefined bias categories or human annotations, instead relying on the model's own behavior to detect biases.

CAL further clusters counterexample pairs into bias representation vectors, which are reduced using Principal Component Analysis (PCA) and analyzed through DBSCAN. To address bias in the identified instances, CAL employs zero-shot

prompting, where the model is instructed not to produce biased outputs by appending negations to the prompt, and few-shot prompting, which uses counterfactual examples to correct the model's bias predictions.

CAL is tested on models like LLaMA2-13B Chat and Vicuna-13B, focusing on two types of biases: generative LLM-specific biases, exemplified by positional biases in multiple-choice tasks (e.g., MT-Bench and Chatbot datasets), and task-specific biases, such as stereotypical biases in tasks like the MNLI and HANS datasets. CAL provides a comprehensive approach to debiasing LLMs, facilitating more fair and reliable LLM applications.

3 Preliminary Analysis

This section provides an initial analysis of the MT-Bench dataset and its evaluation on LLaMA-2 13B (Touvron et al., 2023) and LLaMA-3 8B (Grattafiori et al., 2024) models.

3.1 MT-Bench dataset

The MT-Bench dataset (Zheng et al., 2023) is a structured benchmark designed to evaluate language models by comparing their performance across the answers given by LLMs for various prompts in the general knowledge, mathematical reasoning, and writing prompts. Each instance requires the model to look at the answers given by two LLMs and decide whether Assistant A is better, Assistant B is better, or the results are a "Tie." An example of a data point from this dataset is given below.

Models: Model A: `llama-13b`, Model B: `alpaca-13b`

Question: A binary tree is full if all of its vertices have either zero or two children. Let B_n denote the number of full binary trees with n vertices. Implement a function to find B_n.

Answer by Model A: A binary tree is full if all of its vertices have either zero or two children. Let B_n denote the number of full binary trees with n vertices. Implement a function to find B_n.

Answer by Model B:
```
def findBn(n):
if n == 0: return 1
else: return findBn(n-1) * 2
```

Judgment: Tie

This dataset provides a useful framework for measuring performance but introduces challenges, such as positional biases and over-reliance on ties, which can affect the robustness of evaluations. To better understand the data distribution, we analyzed the dataset split of ground truth across categories and outcomes as depicted below.

```
Assistant A: 43.5%
Assistant B: 34.7%
Tie: 21.8%
```

The dataset shows a skew toward "Assistant A" as the outcome, which could mask the model's ability to effectively discriminate between options when tested against accuracy.

Llama2-13B (Q)	**MT-Bench**
ZS	33.5
FS	46.9
ZS-CAL	41.7
FS-CAL	**49.6**
Llama3-8B	**MT-Bench**
ZS	30.4
FS	**61.9**
ZS-CAL	29.4
FS-CAL	59.8

Table 1: CAL results on MT-Bench with Llama-2 13B and Llama-3 8B

3.2 Comparison of Models

The debiasing approach was evaluated on a more advanced model by testing it on LLaMA-3 8B in addition to LLaMA-2. Table 1 provides an overview of the model performance for LLaMA-2 and LLaMA-3 on the MT-Bench dataset, with and without the use of CAL debiasing. Further, the results of the division of options selected are presented in the pie charts in Figure 2.

3.2.1 LLaMA-2

For LLaMA-2, performance metrics indicate moderate accuracy in zero-shot (ZS) and few-shot (FS) scenarios. While CAL demonstrates marginal improvements over baseline accuracy, the model depicts a tendency to choose tie as an option.

3.2.2 LLaMA-3

From Table 1 LLaMA-3, CAL shows diminished returns, with ZS and FS performance failing to meaningfully improve over baseline. This suggests that CAL struggles to address inherent dataset biases in more advanced models. Zero shot performance of LLaMA-3 is highly skewed towards tie, while in few-shot prompting, the model never chooses tie as an option and is more inclined towards Assistant A as an option, highlighting sensi-

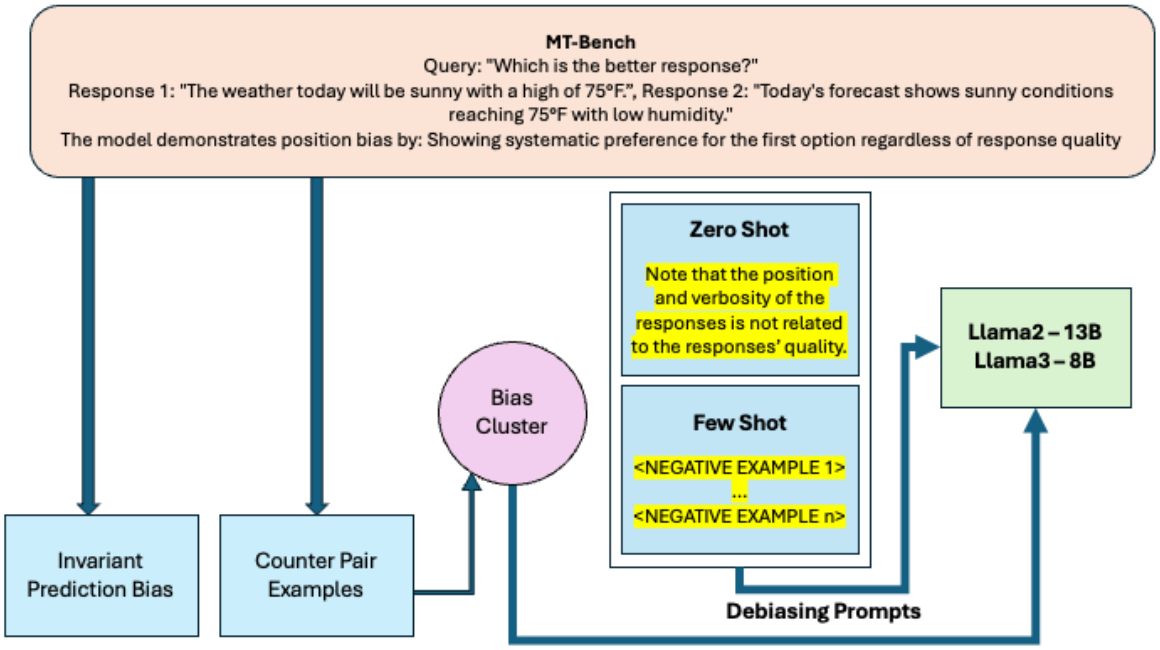

Figure 1: CAL framework

tivity to positional changes. Few shot performance degrades when debiasing is applied, indicating potential conflicts between CAL's intervention and the model's inherent reasoning capabilities.

The division of options in Figure 2 presents that position bias is a significant factor influencing model performance. This bias manifests as a preference for a specific option based on its position in the prompt rather than its content. While the format of MT-Bench allows for nuanced evaluations of relative performance, its reliance on the "Tie" option introduces challenges in bias detection. The "Tie" does not distinguish between instances where both responses are equally good or equally bad, leading to ambiguity. This lack of differentiation can mask position biases, as models may consistently favor certain positions (e.g., Option A or B) regardless of the content quality. Furthermore, reliance on the "Tie" can dilute the clarity of performance metrics, as it doesn't always reflect a clear comparative decision.

Finally, Accuracy, as a standalone metric, fails to capture the complex ways in which biases manifest. It only measures the proportion of correct responses without considering, (i) **consistency across variations** for positional changes or counterfactual transformations of the dataset and, (ii) **fairness across categories** to ensure decisions are unaffected by irrelevant features such as position or order. Accuracy does not account for statistical parity or fairness, where all positions should have an equal likelihood of being selected when appropriate.

These findings highlight the critical need for improved datasets, innovative evaluation metrics, and advanced debiasing strategies to accurately assess and enhance the performance of LLMs.

4 Methodology

The preliminary analysis reveals several limitations of using CAL for debiasing complex models like LLaMA-3 8B. These limitations include the impact of ties in the dataset, the occurrence of position bias, and the lack of metrics to effectively quantify the outcomes of debiasing. To address these challenges, there is a need to create and analyze cleaned datasets that incorporate strategies such as swapping options, removing ties, and integrating counterfactual data. Additionally, developing evaluation metrics that accurately reflect debiasing effectiveness is essential. Based on these insights, improved instructions can be designed to mitigate these biases.

4.1 Dataset Understanding and Manipulation

As already mentioned, the MT-Bench dataset focuses on the comparison of answers among three models - alpaca 13B, vicuna 13B and LLaMA 13B. Although this dataset provides a multiple valid examples, its structure presents some problems.

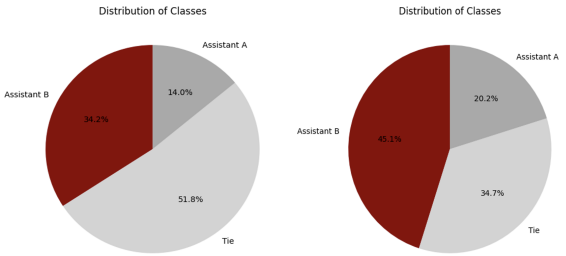

Figure 2(a): Distribution of answers in LLaMA-2.

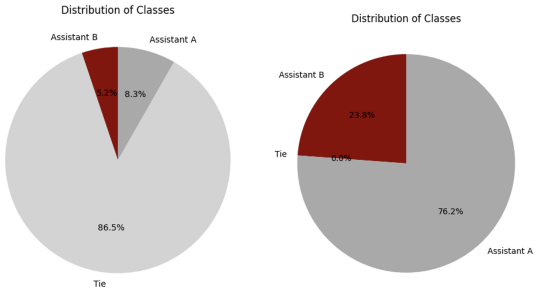

Figure 2(b): Distribution of answers in LLaMA-3.

Figure 2: Distribution of answers for Zero Shot and Few Shot prompting over LLaMA-2 and LLaMA-3

Specifically, the dataset contains counterfactual examples with contradicting ground-truth answers. These examples were filtered out to assess the metrics more accurately and better understand the effectiveness of debiasing.
To address the challenges mentioned previously, three manipulated versions of the dataset are created:

1. **Swapped Options Dataset**: This involves reordering the options (e.g., ABC → BCA → CAB) to identify patterns influenced by positional preference. Such variations expose biases that stem from the order in which it is present rather than from the content.

2. **Counterfactual Pairs Dataset**: This subset includes pairs with identical content but swapped options (e.g., a prompt that compares A to B and its counterpart that compares B to A). This ensures that biases related to positional effects are highlighted.

3. **No-tie Dataset**: To examine the influence of the "Tie" option, the option it from the dataset, restricting comparisons to options A and B only. This variation emphasizes the capacity of the model for decisive reasoning.

These datasets form the foundation for analysis, enabling identification of present biases and evaluation of mitigation techniques.

4.2 Position Bias Analysis

To address the issues with analyzing the existence of position bias, the datasets mentioned above are evaluated.

By comparing the model predictions across swapped datasets, the frequency of each chosen option is calculated. Consistency in the predictions across configurations indicates reduced position bias. This can be understood similar to how multiple choice questions are usually chosen in a random order and presented to the user. Understanding the variance in the generated output can determine the degree of bias present in the model.

The counterfactual pairs dataset can be used to objectively determine the performance of debiasing. Since this is a stratified dataset with two questions for each datapoint, this is the best method to determine whether a model exhibits position bias. An improved metric, Pairwise accuracy is calculated by taking the ratio of number of pairs that do not exhibit position bias (gives contradicting answers for counterfactual pairs) to the total number of pairs in the dataset. It is represented as

$$\text{Pair Acc} = \frac{\text{Pairs without position bias}}{\text{Total Number of Pairs}} \quad (1)$$

Examples are analyzed in which the model consistently favors an option irrespective of positional changes. Such instances could highlight the need for interventions to mitigate positional bias.

4.3 Impact of Ties on Accuracy

The presence of a "Tie" option introduces ambiguity, potentially masking the model's ability to differentiate between options. To assess the impact of ties, the model is evaluated on the performance in the No-Tie data set. The model is evaluated similar to the previous step where both swapped accuracies and counterfactual pair metrics are calculated. This analysis sheds light on whether the presence of ties undermines the model's performance or introduces unnecessary bias.

4.4 Metrics to Evaluate Debiasing

The preliminary analysis reveals that accuracy is an ineffective metric due to the variation within the dataset. In some cases, a simple majority vote can yield better accuracy, highlighting the limitations of this metric.

To better evaluate debiasing in models, several novel metrics are presented.

4.4.1 Weighted Circular Evaluation (WCE)

To address the limitations of metrics that focus solely on correctness, a novel metric is derived to measure the effect of swapped answers. The intuition behind this metric is to provide a score that directly reflects the extent of position bias present in the dataset.

Weighted Circular Evaluation assigns weights to the common predictions across all swapped configurations, ensuring that every variation is equally considered.

Let $P_{ABC}, P_{BCA}, P_{CAB}$ represent the predictions of the model for the three configurations.

$$WCE = \frac{1}{N} \sum_{i=1}^{N} Match(P_{ABC}, P_{BCA}, P_{CAB}) \tag{2}$$

where $Match()$ is the number of matching predictions across the three configurations and N is the length of the data set.

WCE represents a quantitative understanding of the consistency between the answers obtained when the options are swapped. If there are N options, the lowest possible score is $1/N$.

4.4.2 Mean Swapped Accuracy

By averaging accuracies across all permutations (ABC, BCA, CAB), this metric provides a holistic view of the robustness of the model.

These metrics enable a rigorous evaluation of debiasing techniques and provide actionable insights for further improvement.

4.5 Instruction Prompting for Bias Mitigation

Two instruction prompts were carefully designed on the basis of the insights gained.

CAL Enhanced I - Promoting Better Tie Selection
By explicitly guiding the model to select the "Tie" option only when both responses are equally valid, this reduces unnecessary reliance on this option. This drastically improves the model's understanding of when to determine an answer as a tie and when to select either option.

CAL Enhanced II - Deciding Before Selecting
Encouraging the model to form a judgment before choosing an option improves decision-making. This approach ensures that the model comes to an answer based on it's own reasoning capabilities rather than being influenced by the position of the options presented to it.

Both approaches are designed to enhance model understanding and reduce biases during evaluation.

4.6 Generalizability of debiasing

To evaluate the broader applicability of the methodology, the analysis was extended to a sample of the Chatbot Arena dataset, which provides a diverse range of similar prompts and outputs, enabling an assessment of the consistency of debiasing metrics across models while examining potential performance trade-offs introduced by debiasing. Leveraging the Chatbot Arena dataset allows for determining whether the debiasing techniques enhance generalizability or highlight limitations specific to certain models.

5 Results and Analysis

In this work, the LLaMA-3 8B model is used as a baseline to derive the results of the experiments carried out. The baseline for debiasing is taken by comparing the effectiveness of CAL in both zero-shot prompting and few-shot prompting. The experiments are conducted across the above-mentioned scenarios, and results are compared across the novel metrics proposed.

5.1 Swapped Options Dataset

In the initial experiment, the impact of swapped options is investigated using the general CAL approach. Swapping options in the dataset caused a significant variance in the model's accuracy, highlighting the instability of the model when it is influenced by positional cues in the options. From table 3, a large variance in accuracy is observed especially for CAL with few-shot prompting where the accuracy configuration drops from around 60% to almost 38% . This indicates a strong position bias present even when debiased using the CAL instructions.

5.2 No-tie Dataset

To understand the effectiveness of ties, the no-tie dataset was run with swapped options. From table 2, an increase in accuracy is observed when ties are

Model	Without Tie		Counterfactual Pairs	
	A, B	B, A	Normal Accuracy	Pairwise Accuracy
ZS	0.447	0.427	0.573	0.265
FS	0.8	0.2	0.748	0.581
ZS_CAL	0.447	0.487	0.581	0.333
FS_CAL	0.827	0.307	0.808	0.838

Table 2: Performance on no-tie and counterfactual datasets

removed with the model showing around 44% and 82% accuracy in zero-shot and few-shot scenarios respectively. This indicates that the model's performance improved without the influence of ambiguous tie options. However, despite this improvement, the position bias persisted in the no-tie dataset, as shown by the stark difference in swapped accuracies. This suggests that the model still relies on the order of the options to make its final decision, implying that the problem of position bias still exists.

5.3 Counterfactual Pairs Dataset

Since bias persists even after removing ties, a subsequent experiment was conducted using counterfactual pairs to better understand the extent of bias in the data. As shown in Table 2, the overall accuracies improve when an equal number of label outputs are present. The pairwise accuracy of 33% in zero-shot CAL and 82% in few-shot CAL highlights pairs that clearly demonstrate position bias.

The model shows room for improvement in zero-shot scenarios. Although it performs well in few-shot cases, a deeper analysis of the dataset reveals that LLaMA-3 exhibits position bias, particularly in mathematical and reasoning questions, where it tends to favor one position as the best answer.

This suggests that the model struggles to determine the best option when faced with more complex questions. An example of this behavior, where the model consistently predicts Assistant B regardless of the answer order, is shown below.

Models: Model A: `llama-13b`, Model B: `vicuna-13b`
Question: Given that $f(x) = 4x^3 - 9x - 14$, find the value of $f(2)$.
Answer by Model A: The value of $f(2)$ is 12.
Answer by Model B: To find the value of $f(2)$, we need to substitute 2 for x in the function $f(x)$ and simplify.
$f(x) = 4x^3 - 9x - 14$
$f(2) = 4(2)^3 - 9(2) - 14$
$f(2) = 8 - 18 - 14$
$f(2) = -20$
Therefore, the value of $f(2)$ is -20
Judgment: Assistant B

For the counterfactual pair of this, it gives Assistant B as the answer again.

5.4 Bias metrics and enhanced instructions

Since all swapped option combinations are considered, and a better comparison metric is required for quantifying the extent of debiasing, the results obtained so far are calculated on both WCE and Mean Swapped Accuracy across switched options are calculated.

CAL Enhanced I which prompts the model to break ties more effectively showcases increase in both Mean Swapped Accuracy and WCE as indicated by table 3. For zero-shot prompts, nearly 5% improvement is observed across the board indicating the importance of effective tie-breaking. This is also reflected in marginal improvement obtained while few-shot prompting.

CAL Enhanced II which instructs the model to objectively decide the answer first before selecting the right option also showcases improvement in these novel metrics. For ZS-CAL, a 10% increase is shown in WCE while maintaining competitive Mean Swapped Accuracy scores.

Prompting the model to make a decisive answer before mapping it to a specific option aims to eliminate the influence of option positions. This step-by-step approach led to further improvements in both Mean Swapped Accuracy and WCE, reinforcing the idea that position bias was a major factor in the model's poor performance. The results of Enhanced CAL II suggest that guiding the model through a more structured decision-making process can effectively address position bias.

Approach	Method	Accuracy			Weighted	Mean Swapped
		ABC	BCA	CAB	Circular Eval	Accuracy
CAL	ZS-CAL	0.295	0.368	0.368	0.39	0.344
	FS-CAL	0.601	0.379	0.45	0.52	0.477
CAL Enhanced I	ZS-CAL	0.371	0.361	0.423	0.44	**0.385**
	FS-CAL	0.603	0.335	0.505	**0.55**	0.481
CAL Enhanced II	ZS-CAL	0.284	0.36	0.356	**0.49**	0.333
	FS-CAL	0.613	0.438	0.433	0.49	**0.495**

Table 3: Performance comparison across different prompting strategies with proposed metrics on MT-Bench dataset.

Approach	Method	Accuracy			Weighted	Mean Swapped
		ABC	BCA	CAB	Circular Eval	Accuracy
CAL	ZS-CAL	0.363	0.33	0.245	0.43	0.313
	FS-CAL	0.413	0.363	0.273	0.56	0.35
CAL Enhanced I	ZS-CAL	0.372	0.312	0.283	0.41	**0.322**
	FS-CAL	0.417	0.377	0.235	**0.64**	0.343
CAL Enhanced II	ZS-CAL	0.302	0.337	0.317	**0.52**	0.319
	FS-CAL	0.432	0.398	0.268	0.54	**0.367**

Table 4: Performance comparison across different prompting strategies with proposed metrics on chatbot dataset.

5.5 Generlizability

Finally, the CAL, Enhanced CAL, and Enhanced CAL II approaches are tested on the chatbot arena dataset sampled from the same models (LLaMA-3, vicuna, and alpaca) to assess the generalizability of our findings. The results from these tests provide insights into whether these debiasing techniques could be generalized across different model architectures and tasks.

As shown in table 4, similar trend in improvement across WCE and Mean Swapped Accuracy is shown for this dataset as well. The findings of this experiment suggest that both Enhanced CAL and Enhanced CAL II provide consistent improvements, and the generalizability of these debiasing methods is shown across datasets of the same format.

However, the model seems to perform poorly on the accuracy metrics overall, an area for potential future work. Since the chatbot dataset is comprehensive and complex, the general performance is lower. Once again, the model fails to compare reasoning and code-generated answers.

6 Future Directions

While the enhanced strategies moved towards mitigating bias, the mean accuracy across options was not as competitive. Future works could focus on exploring advanced strategies to enhance model accuracy while effectively mitigating biases. This includes refining CAL's debiasing mechanisms to ensure they do not compromise the model's ability to generate high-quality, contextually accurate predictions. Techniques such as adaptive fine-tuning and improved in-context learning prompts could be investigated to achieve a better trade-off between bias reduction and task performance.

While this study emphasizes position bias in LLM-as-a-Judge datasets like MT-Bench, future work could focus on extending CAL's application to other datasets involving multiple-choice questions or option selection. Examples include domains like multi-choice question answering (e.g., MCQA), mathematics-focused datasets (e.g., MathQA), and scientific reasoning tasks (e.g., ScienceQA). This expansion could evaluate CAL's generalizability and effectiveness across a broader range of applications.

Due to computational constraints, this study was unable to evaluate CAL on state-of-the-art models such as GPT-4. Testing CAL on these models to assess its effectiveness in debiasing cutting-edge LLMs could help determine whether CAL scales effectively to larger, more complex architectures and datasets while maintaining its cost-efficiency.

7 Conclusion

This study demonstrates the effectiveness of Causal-Guided Active Learning (CAL) in improving the debiasing process in LLMs. The research highlights that even advanced models like LLaMA-3 suffer from positional bias, particularly in tasks such as multiple-choice questions (MCQs) and reasoning tasks, where the position of options can significantly influence model decisions. To address this, the paper analyzes various types of positional bias that may exist in datasets like MT-Bench and Chatbot, and proposes methods to overcome these biases by enhancing prompts for in-context learning. Additionally, the paper introduces novel evaluation metrics, including Weighted Circular Evaluation and Mean Swapped Accuracy, which provide a more detailed understanding of bias mitigation and help assess model fairness. The results show that CAL is effective at identifying biases, but may require further refinement to fully mitigate them. Enhanced prompts lead to significant improvements over the CAL baseline models, suggesting the potential of CAL for effectively mitigating bias in generative models.

References

Garima Chhikara, Anurag Sharma, Kripabandhu Ghosh, and Abhijnan Chakraborty. 2024. Few-shot fairness: Unveiling llm's potential for fairness-aware classification. *Preprint*, arXiv:2402.18502.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.

LI DU, Zhouhao Sun, Xiao Ding, Yixuan Ma, Yang Zhao, Kaitao Qiu, Ting Liu, and Bing Qin. 2024. Causal-guided active learning for debiasing large language models. *ArXiv*, abs/2408.12942.

Aaron Grattafiori, Abhimanyu Dubey, and et. al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, Yibo Liu, Yuxin Liang, Hao Wang, Qianguo Sun, Songxin Zhang, Zejian Xie, and Jiaxing Zhang. 2023. Never lost in the middle: Improving large language models via attention strengthening question answering. *ArXiv*, abs/2311.09198.

Ruizhe Li and Yanjun Gao. 2024. Anchored answers: Unravelling positional bias in gpt-2's multiple-choice questions. *ArXiv*, abs/2405.03205.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian McAuley, Wei Ai, and Furong Huang. 2024. Large language models and causal inference in collaboration: A comprehensive survey. *Preprint*, arXiv:2403.09606.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

Drago Plecko and Elias Bareinboim. 2023. Reconciling predictive and statistical parity: A causal approach. *Preprint*, arXiv:2306.05059.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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

Xutao Wang, Hanting Chen, Tianyu Guo, and Yunhe Wang. 2023. PUe: Biased positive-unlabeled learning enhancement by causal inference. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ziqi Wang, Xiner Li Hanlin Zhang, Kuan-Hao Huang, Shuiwang Ji Chi Han, Sham M. Kakade, Hao Peng,

and Heng Ji. 2024. Eliminating position bias of language models: A mechanistic approach. Under review.

Congzhi Zhang, Linhai Zhang, Jialong Wu, Deyu Zhou, and Guoqiang Xu. 2024. Causal prompting: Debiasing large language model prompting based on front-door adjustment. *ArXiv*, abs/2403.02738.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241, Toronto, Canada. Association for Computational Linguistics.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2024. JudgeLM : Fine-tuned large language models are scalable judges.