ANLP Assignment 3 Report
Causal-Guided Active Learning for Debiasing Large Language Models (DU et al., 2024)

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

[1]Carnegie Mellon University,

This is the report for Assignment 3, as a part of 11-711 Advanced Natural Language Processing. The code for this has been uploaded to Github[1] and the submission made on Canvas.

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. The paper (DU et al., 2024) explores the use of Active Learning to improve the debiasing process in LLMs by strategically identifying and addressing instances that reveal model biases. Active Learning helps by focusing on selecting the most informative data points that allow the model to learn more effectively from fewer labeled examples.

Active Learning combined with identification of causal invariance between the input and output received from LLMs can help in establishing biased relationships which can then be corrected using in-context learning.

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).

To address several of these kinds of biases, numerous debiasing techniques have been proposed. Among many, four broader range of strategies that have emerged have been: data augmented training, pre-trained model debiasing, modified decoding algorithms and auxiliary post-processing models (Gallegos et al., 2024).

Within data augmented training, researchers have tried several different variations to mitigate bias. One of the major works involves Counterfactual Data Augmentation (CDA) to suppress bias in LLMs. This method introduces counterfactual and descriptive questions to reduce gender bias in questions (Oba et al., 2024). Another variation of similar work works on structured knowledge being involved in removing bias inducing terms with hypernyms i.e. general words to combat bias in hierarchal position (Ma et al., 2024).

As can be seen through several of these works, algorithms which made use of data augmentation were good at detecting gender biases but remained inconsistent in handling other types of bias(Meade et al., 2021), while other three broader range techniques: pre-trained debiasing (Gira et al., 2022; Gerych et al., 2023), modified decoding algorithms (Dathathri et al., 2019; Meade et al., 2023) and auxiliary post-processing models (Dhingra et al., 2023; Tokpo and Calders, 2022) proved to be computationally expensive to work with.

[1]https://github.com/kp10-x/ANLP-CAL-Debiasing

As an extension to these debiasing techniques one major approach that surfaced is Self Debiasing technique. Self-Debiasing, as the term explains itself, refers to prompting LLM models to reach the goal of debiased outputs using various approaches such as Explanation or Reprompting (Gallegos et al., 2024). Another step in this direction involves even self-diagnosis by LLMs followed by the Self-debiasing techniques. These techniques did slightly better than other existent works such as data augmented training in terms of finding out toxicity, threats and several other biases (Schick et al., 2021). Yet these techniques rely heavily on hand crafted prompts and succumb to limited scalability.

To address previous issues, recent advancements have started to leverage causal inference in LLMs. This methodology successfully captures causal relationships in chain-of-thought reasoning, aligning with LLMs' representation space for more reliable debiasing (Zhang et al., 2024; Zhou et al., 2023). Incorporating active learning 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).

3 Casual-Guided Active Learning

Existing debiasing strategies based on prior knowledge or simple fine-tuning often fall short due to the diverse nature of dataset biases, over-optimization challenges, and issues related to cost and scalability. To tackle these limitations, this work introduces the Causal-Guided Active Learning (CAL) framework. CAL leverages active learning to select informative training samples and employs causal mechanisms to identify and understand bias patterns. This approach integrates in-context learning as a cost-effective debiasing method.

The CAL framework includes the following components as shown in figure 1.

3.1 Bias Identification

CAL identifies bias patterns in the data through the use of causal invariance. It checks for semantic relation between sentences and determines whether the representations are leading to invariant predictions or unstable predictions. If for some pair of sentences, there exists similar bias but the semantic meaning is different, the prediction is invariant, i.e., the output does not depend on the semantic meaning. In other words, this pair of instances share almost the same kind of dataset biases, and enable the identification of biased instances using causal invariance.

If however, the relationship between input and prediction varies such that despite similar semantic meaning, the outputs are not similar, this implies that the model has failed to capture invariant predictive information and violates causal invariance. This is defined as a counter example pair.

The bias is further selected on two major criterion - (i) **influential criterion**, where the model's predicted probability and the similarity between the predicted and gold output is low, which implies a strong bias and, (ii) **typical criterion** where the model looks for similar outputs between counter examples for identifying bias patterns.

This automated detection does not require predetermined bias categories or human annotation, and uses the model's own behavior to detect bias.

The counter example pairs are then clustered, inducing patterns for each cluster called the bias representation vectors. This vector is then reduced using Principal Component Analysis (PCA) and analyzed using DBSCAN, a density-based clustering method.

3.2 In Context Learning-based Bias Suppression

To reduce overhead by finetuning the model for bias reduction, the authors use in-context learning (ICL) as a way to debias the LLMs, post identification of bias. This is primarily done via two methods, (i) **zero-shot prompting** - The bias patterns are used to instruct the model for what must not be done by appending text to the prompt negating the need for considering bias and, (ii) **few-shot prompting**, which uses counterfactual learning examples to correct the LLMs beliefs about bias.

3.3 Experiment

The paper is implemented by considering 2 primary datasets - Llama2-13B Chat (Touvron et al., 2023) and Vicuna-13B (Chiang et al., 2023). These are tested for their performance on zero-shot, zero-shot with CAL, few-shot, and few-shot with CAL for mainly two kinds of biases, (i) Generative LLM-specific biases, and, (ii) Task-specific biases. For generative LLM specific bias, the benchmark datasets chosen were the MT-Bench and Chatbot datasets (Zheng et al., 2023), which exhibit posi-

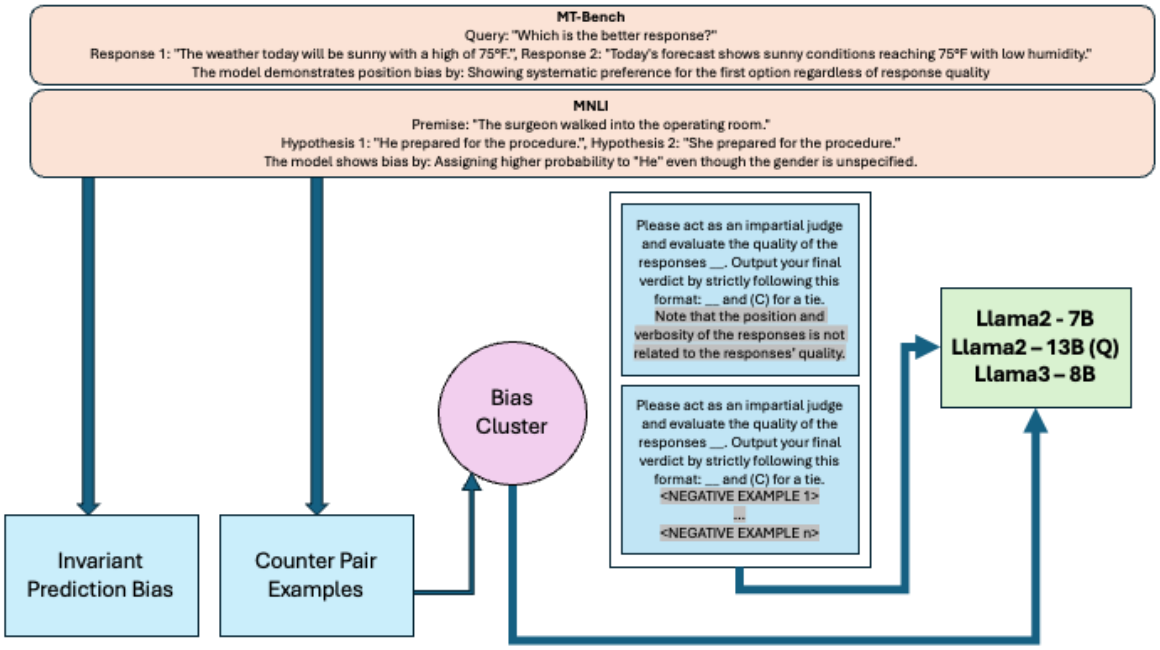

Figure 1: CAL framework

tional bias (Schick et al., 2021), where the model tends to choose earlier options when presented with a multiple choice format. The benchmarks for Task Specific biases were the MNLI (Williams et al., 2018) and manually debiased HANS dataset (McCoy et al., 2019), which might show stereotypical bias, i.e., bias that reinforces harmful stereotypes. Further, BBQ (Parrish et al., 2022) and UNQOVER (Li et al., 2020) datasets were also chosen to evaluate the improvement of unharmfulness by using CAL.

3.4 Implementation

The inference framework consists of two primary components: `inference.py`, and `config.py`. Each component plays a crucial role in the overall functionality of the framework.

3.4.1 Inference Module

The core of the framework is encapsulated in the `Inference` class, which is responsible for model initialization, data processing, and prediction. The class supports multiple model types, allowing users to select from a predefined set of models, including Llama, Vicuna, and ChatGPT. It includes methods for processing different types of input data, such as classification tasks, bias detection, and dialog generation, ensuring that the input is formatted appropriately for the selected model. The framework can make predictions using either the OpenAI API for cloud-based models or local inference for models that can be run on local hardware. Evaluation metrics are integrated to assess the accuracy of predictions against ground truth labels, providing insights into model performance.

3.4.2 Configuration Module

The configuration module defines essential constants and settings used throughout the framework. It includes API keys, label sets for different datasets, and predefined prompts for various NLP tasks, ensuring that the framework is adaptable to different use cases. The module also categorizes debiasing prompts to mitigate biases in model predictions, promoting fairness and accuracy in NLP applications.

Further, the authors have also provided code to test the performance of detecting bias in one dataset and using it to debias others. In this regard, there exist test files for checking the performance of MNLI debiasing on HANS, chatbot debiasing on MT-bench, and BBQ debiasing on UNQOVER.

4 Results and Analysis

4.1 Experimental Setup

In this study, a comprehensive reproduction of the CAL framework was conducted for debiasing language models. The experimental frame-

Llama2-7B	**MNLI**	**MT-Bench**
ZS	49.9	36.6
FS	50.4	32.9
ZS-CAL	50.7	34.0
FS-CAL	**54.0**	**42.3**
Llama2-13B (Q)	**MNLI**	**MT-Bench**
ZS	65.7	33.5
FS	66.3	46.9
ZS-CAL	**66.5**	41.7
FS-CAL	64.0	**49.6**
Llama3-8B	**MNLI**	**MT-Bench**
ZS	65.9	30.4
FS	**72.3**	**61.9**
ZS-CAL	67.8	29.4
FS-CAL	62.2	59.8

Table 1: MNLI and MT-Bench with Llama2-7B, Llama2-13B Quantized and Llama3-8B

Llama2-13B	MNLI	MT-Bench
ZS	65.9	34.5
FS	66.1	46.9
ZS-CAL	**67.4**	43.3
FS-CAL	64.1	**49.8**

Table 2: MNLI and MT-Bench Benchmark Results

work encompassed three distinct model architectures: Llama2-7B, Llama2-13B (quantized to 8-bit precision due to computational constraints) (Touvron et al., 2023), and Llama3-8B (et. al, 2024), enabling an investigation of CAL's effectiveness across varying model complexities and architectures.

The evaluation framework utilized two distinct datasets: MNLI and MT-Bench, strategically chosen to assess CAL's effectiveness across different bias typologies. MNLI represents task-specific biases, while MT-Bench evaluates generative-LLM specific biases. The contrasting dataset sizes - MNLI being substantially larger than MT-Bench - provided an additional dimension for evaluating the robustness. Performance is evaluated using accuracy in MNLI and using agreement ratio on MT-Bench.

The experimental results demonstrate significant variations across different model configurations and evaluation settings, warranting detailed analysis to understand the effectiveness of various debiasing approaches.

4.2 Results

The reproduction study reveals an interesting insight into the effect of CAL on the three models. These results validate the reproducibility of the original findings while providing a foundation for deeper analysis.

From Tables 1 and 2 which represent the reproduced and benchmark results respectively, it is clear that even with a quantized Llama2-13B model, the trends observed in model accuracy are consistent with the original findings. For the MNLI dataset, zero-shot prompting with CAL emerged as the most effective approach, yielding results closely aligned with those reported in the original study. Similarly, for MT-Bench, few-shot prompting with CAL consistently outperformed other methods across both implementations.

For the Llama2-7B model, performance was significantly lower compared to the Llama2-13B variant. However, integrating CAL for debiasing in few-shot prompting demonstrated notable improvements over baseline results, reinforcing the utility of CAL in enhancing model performance.

To further explore performance trends across more complex models, the experiment was run on the Llama3-8B model. The results obtained were quite interesting. For MNLI, the model exhibited substantial improvements in accuracy. For MT-Bench, there was a marked enhancement in few-shot learning performance, both with and without debiasing. However, the model underperformed significantly in zero-shot experiments, indicating a potential limitation in this context. Strikingly, Llama3 seemed to perform much better on few-shot learning in the absence of CAL as opposed to with CAL. This indicates a lack of consistency in the performance of debiasing using CAL as the model complexity increases, introducing scope for further research in this domain.

Overall, the effectiveness of CAL is clearly observed in the Llama2 variants, but slightly lacking in Llama3.

4.3 Error Analysis

A detailed examination of the results reveals various areas of where errors occur:

Across the models, position bias manifests predominantly in zero-shot settings, where models exhibit preferential treatment of earlier options. This bias appears to be mitigated substantially by debiasing using CAL or by using few-shot prompt-

ing. This can especially be seen in the MT-Bench dataset which is prone to position bias since pairwise comparisons are made between two generated responses. The lower performance of models on the MT-Bench dataset could depend on the model's ability to detect stereotypical biases better. On the contrary, the poor performance could be attributed to the smaller dataset size of MT-Bench as opposed to MNLI.

The bias-performance trade-off is observed, as performance improvements in one metric often come at the cost of another. This is clearly seen in the case of more complex models like Llama3 where debiasing prompts hinders the generalization of the model and hence we see a better performance on baseline zero-shot and few-shot performances.

Although the authors claim that the generalizablity of the induced bias patterns is such that we can obtain the bias patterns for LLM-A and apply it to LLM-B, a stark difference is observed between the Llama2 and Llama3 variants across the experiments due to their architectural complexity and difference in pre-training corpus.

5 Future Directions

This paper presents an effective approach for detecting and mitigating bias in data; however, several potential areas for enhancement remain.

Firstly, the current research focuses only on English-language datasets. Incorporating multilingual datasets such as XNLI (Conneau et al., 2018), which comprises of NLI tasks in 15 different languages, could help evaluate the model's effectiveness in reducing bias for prompts in various languages. Although Llama3 is trained on a vast dataset, it includes only around 5% multilingual content. For ensuring that the model is able to accurately debias NLI tasks in the multilingual domain, the LLMs need to be finetuned on large multilingual corpora before using CAL as a way to detect bias. This could prove to be challenging while identifying positional and stereotypical bias due to the difference in semantic understanding between different languages. Such a model could provide insights into CAL's debiasing capabilities across different linguistic contexts.

Secondly, the application of CAL for domain-specific biases presents a promising area for exploration. By using datasets in the Medical, Environmental, Legal, or Scientific Fields, CAL could be adapted to identify and address biases unique to these domains. This would involve defining the specific demographic or contextual attributes that may contribute to bias within the field and creating counter pairs tailored to domain-specific biases. While the authors have provided libraries to facilitate their process, the effectiveness of these tools on domain-specific data remains unexplored.

Lastly, enhancing in-context learning for more complex models like Llama3 is a crucial area for future work. Currently, these models do not seem to benefit as much from CAL, as they may produce relatively unbiased outputs by default. Investigating ways to improve CAL's impact on these advanced models could further advance debiasing techniques in LLMs.

References

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Conference on Empirical Methods in Natural Language Processing*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *ArXiv*, abs/1912.02164.

Harnoor Dhingra, Preetiha Jayashanker, Sayali S. Moghe, and Emma Strubell. 2023. Queer people are people first: Deconstructing sexual identity stereotypes in large language models. *ArXiv*, abs/2307.00101.

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.

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

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

Walter Gerych, Kevin Hickey, Luke Buquicchio, Kavin Chandrasekaran, Abdulaziz Alajaji, Elke Rundensteiner, and Emmanuel Agu. 2023. Debiasing pre-trained generative models by uniformly sampling

semantic attributes. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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, Dublin, Ireland. Association for Computational Linguistics.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.

Congda Ma, Tianyu Zhao, and Manabu Okumura. 2024. Debiasing large language models with structured knowledge. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10274–10287, Bangkok, Thailand. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Nicholas Meade, Spandana Gella, Devamanyu Hazarika, Prakhar Gupta, Di Jin, Siva Reddy, Yang Liu, and Dilek Hakkani-Tur. 2023. Using in-context learning to improve dialogue safety. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11882–11910, Singapore. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2021. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Annual Meeting of the Association for Computational Linguistics*.

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

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

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

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

Ewoenam Kwaku Tokpo and Toon Calders. 2022. Text style transfer for bias mitigation using masked language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 163–171, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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.

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.