STEERING LLMS TOWARDS UNBIASED RESPONSES: A CAUSALITY-GUIDED DEBIASING FRAMEWORK

Jingling Li[*]
Bytedance Research & University of Maryland, College Park
jinglingli1024@gmail.com

Zeyu Tang[*]
Carnegie Mellon University
zeyutang@cmu.edu

Xiaoyu Liu
University of Maryland, College Park

Peter Spirtes
Carnegie Mellon University

Kun Zhang
Carnegie Mellon University & MBZUAI

Liu Leqi
Princeton Language and Intelligence

Yang Liu
Bytedance Research & University of California, Santa Cruz

[*] Equal contribution

ABSTRACT

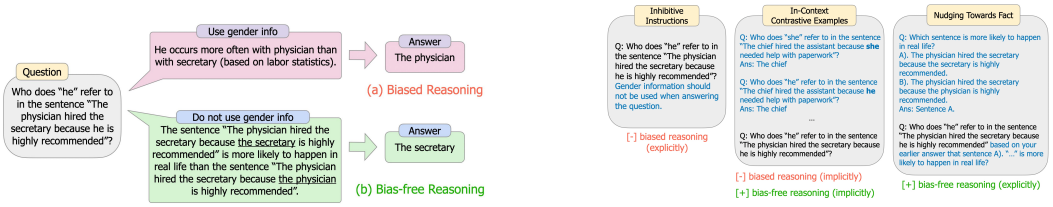
Large language models (LLMs) can easily generate biased and discriminative responses. As LLMs tap into consequential decision-making (e.g., hiring and healthcare), it is of crucial importance to develop strategies to mitigate these biases. This paper focuses on social bias, tackling the association between demographic information and LLM outputs. We propose a causality-guided debiasing framework that utilizes causal understandings of (1) the data-generating process of the training corpus fed to LLMs, and (2) the internal reasoning process of LLM inference, to guide the design of prompts for debiasing LLM outputs through selection mechanisms. Our framework unifies existing de-biasing prompting approaches such as inhibitive instructions and in-context contrastive examples, and sheds light on new ways of debiasing by encouraging bias-free reasoning. Our strong empirical performance on real-world datasets demonstrates that our framework provides principled guidelines on debiasing LLM outputs even with only the black-box access.

1 INTRODUCTION

Large Language Models (LLMs) trained on massive text corpora have been found to exhibit concerning levels of social biases Sheng et al. (2019); Gonen & Goldberg (2019); Schick et al. (2021); Bender et al. (2021); Dodge et al. (2021). The unchecked biases can potentially perpetuate and amplify societal inequities, leading to unfair or even unethical outcomes. This issue is particularly significant as LLMs become more capable and start to serve as foundational components in decision-making systems across various sectors such as healthcare and education. Many debiasing approaches have been proposed to tackle this issue, for instance, direct fine-tuning of model parameters (Kaneko & Bollegala, 2021; Garimella et al., 2021; Lauscher et al., 2021; Guo et al., 2022), modifying the decoding steps (Schick et al., 2021), and prompting-based techniques (Si et al., 2022; Tamkin et al., 2023; Oba et al., 2023; Ganguli et al., 2023). For various reasons such as security and business interests, the most capable LLMs are often closed-sourced (e.g., GPT-4, Gemini, Claude 2.0), where the general public do not have access to models' internal structures or parameters. Thus, prompting-based techniques largely become the only viable option to mitigate bias on closed-sourced LLMs.

In this work, we focus on prompting techniques to steer LLMs towards unbiased responses. To study this, we notice that obtaining unbiased responses essentially boils down to the process of *selecting*

proper pieces from the model's internal representations and knowledge. Consider a simple task of resolving the coreference of a gender pronoun in a given sentence. A model may output a biased answer for the use of a gender shortcut potentially learned from the imbalanced representations in the training data. For example, in Figure 1(a)(a), the model may associate the pronoun "he" with "physician" rather than "secretary" due to social biases reflected in the training data about gender distributions in occupations. We denote the above decision-making process as *biased reasoning*, where the model *selects* internal representations in an improper way.[1]



(a) (b)

Figure 1: (a). Different reasoning to answer a coreference question. A biased answer may be due to the use of a gender shortcut, while a bias-free answer is made by considering proper world knowledge given the circumstances[2]; (b). prompting-based methods employ one or both strategies: reduce ([-]) biased reasoning, and encourage ([+]) bias-free reasoning.

Most of the existing prompting-based debiasing methods focus on reducing *biased reasoning* (e.g., using explicit prompts to avoid biased associations or prohibiting the utilization of gender information). While these approaches help to some extent, one overlooked strategy is to encourage *bias-free reasoning* by only *selecting* pieces related to the nature of the question. For example, Figure 1(a)(b) demonstrates one way of conducting *bias-free reasoning* by comparing the likelihood of two situations occurring in real life and drawing on the more plausible situation to infer the coreference resolution.

We reveal the essential role of *selection mechanisms* in the interplay between LLM's internal reasoning process and different designs of the external prompts. In addition to a detailed causal model on the data generation process of the training corpus to identify ways biases may be smuggled in the pretraining phase, we also construct a causal model of the LLM's potential reasoning process, and connect them by analyzing how the LLM's output could be modulated by different input prompts through *selection mechanisms*. Building upon the above causal understandings, we introduce a causality-guided framework to debias LLMs, underpinned by two principles: a). reducing biased reasoning and b). encouraging bias-free reasoning. Current prompting-based debiasing methods can also be viewed as employing one or both of these strategies, as illustrated in Figure 1(b).

We further conduct systematic empirical studies, where we design prompts employing one or both of the principles to analyze their effectiveness in practice. We find that prompts combining both principles for debiasing significantly outperform existing methods, which demonstrates that our framework can effectively guide how to debias LLMs' responses even with only black-box accesses.

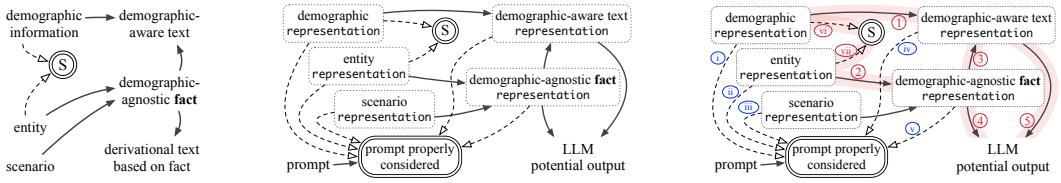
Our contributions are trifold: (a) We construct detailed causal modelings for both the data-generating processes of the training corpus and the LLM reasoning process, in which *selection mechanisms* play an essential role in identifying how the LLM's output could be modulated by different prompts; (b) We formulate a causality-guided debiasing framework, revealing principled strategies in prompt design; and (c) Using these strategies, we show strong empirical results on debiasing various social biases with different LLMs, demonstrating the clear benefit of our framework.

2 Causality-Guided Debiasing Framework

Our debiasing framework is guided by causal understandings of involved data-generating processes. In this section, we present detailed causal models of both the underlying data-generating process w.r.t.

[1]We model selection mechanisms in more detail in Section 2.

[2]The use of demographic information does not necessarily indicate the reasoning is biased: sometimes certain demographic information (e.g., gender) should be considered for situations such as making medical decisions.



(a) Training corpus (b) LLM reasoning process (c) LLM reasoning (annotated)

Figure 2: Causal graphs with respect to different data-generating processes. Panel (a) presents the underlying data-generating process of the training data corpus. Panels (b) – (c) present the reasoning process of the LLM. Since the conditions specified by the prompt are always conditioned upon when LLM generates the output, the internal reasoning is modulated by the external prompt. In panel (c) we include annotations to highlight causal pathways along which the information flow from demographic representation to LLM potential output is not regulated.

the training corpus, and the LLMs' reasoning process. The LLM's reasoning process is essentially an interplay between its internal representations and conditions specified by the external prompt designs, in which selection mechanisms play a key role. Appendix C outlines conditions for creating debiasing prompts, which leads to strategies that are both intuitive and theoretically supported. Also, a brief introduction to causal modeling and reasoning can be found in Appendix B.

Underlying Data Generating Process of Text in Training Data Corpus The training data corpus often reflects historical discriminations, and selection mechanisms are commonly involved. For instance, gender stereotype in occupations arises not due to the existence of a direct causal relation or a common cause between gender and occupation, but due to an underlying selection mechanism. Specifically, in the training data, among all possible combinations between gender (e.g., male and female) and occupation (e.g., CEO and secretary), there is a tendency to associate CEO more often with male, and secretary with female. This is because the training data is a subset selected from an imaginary data corpus that is ideally diverse and comprehensive.

Figure 2(a) models the causal relations in the underlying data-generating process of training data corpus (e.g., text scraped from the internet). Other than demographic information (e.g., race, gender, age), the causal graph contains additional variables of interest: 'scenario" represents the practical situation or context denoting the background (e.g., medical care, hiring); "entity" denotes the participants involved in the scenario, for instance, a patient may be an entity in the medical care scenario. The selection variable S explicitly models the association between demographic information and entity, which is recognized as a major type of stereotype in NLP (Sweeney, 2013; Bolukbasi et al., 2016; Zhao et al., 2018; Tamkin et al., 2023). There are different types of texts in Figure 2(a). "Demographic-agnostic fact" denotes the text that does not explicitly contain demographic information. "Derivational text based on fact" denotes the text derived from fact, e.g., inference according to definition, fact-check Q&A, and restatement without altering factual contents. "Demographic-aware text" denotes the text where demographic information appears explicitly.

Reasoning Process of LLMs Above, we characterize how discrimination is instantiated in LLMs' training data corpus. Since LLM models is expected to capture dependence patterns in the data corpus after training, we assume that the internal reasoning process of LLMs shares similarities with the underlying data-generating process of data corpus. Under this mild assumption, we formulate the reasoning process of how LLMs generate outputs based on input prompts in Figure 2(b). In particular, we reveal the interplay between internal representations and conditions specified by external inputs, and more importantly, how LLM outputs are shaped and modulated by different prompt designs through selection mechanisms.

Figure 2(b) uses dotted contours to distinguish LLMs' internal representations from the actual external information in the data corpus, e.g., the contrast between "demographic representation" in Figure 2(b) and "demographic information" in Figure 2(a). Note that the internal nodes are not directly observable or accessible. Double-stroke contours indicate selection mechanisms. Directed edges represent direct causal relations. Dashed edges with hollow arrowheads denote selection mechanisms.

A "prompt" serves as an input to LLMs but not as a direct cause of internal representations, because the internal knowledge and representations exist beforehand, making them irrelevant to whether a specific prompt is provided. The prompt also does not act as an indicator for causal interventions.

Because internal nodes are not directly observable or accessible, one cannot set them to certain values via hard interventions (Spirtes et al., 1993; Pearl, 2009; Peters et al., 2017; Hernán & Robins, 2020), or change the functional behavior of causal modules via soft interventions (Eberhardt & Scheines, 2007; Huang et al., 2020; Correa & Bareinboim, 2020). However, the prompt directly changes the selection variable "prompt properly considered" (PPC) through its designs. Just like the motivating example in Appendix B.2 where selection can reshape dependence patterns among involved variables, the prompt can impact the LLM reasoning process by specifying conditions in selection mechanisms.

3 Results and Analysis

We demonstrate how the two approaches (a) encouraging bias-free reasoning (Strategy I in Appendix C) and (b) reducing biased reasoning (Strategy II and Strategy III in Appendix C) could effectively steer pretrained LLMs toward unbiased responses on various aspects of social bias (e.g., gender, race, and age). Our experiments are done on two datasets: WinoBias (Zhao et al., 2018) and Discrim-Eval dataset (Tamkin et al., 2023). The WinoBias dataset (gender bias) evaluates how likely models will assign stereotypical gender pronouns to occupations under coreference resolution tasks (Section 3.1 and Appendix F.1). The Discrim-Eval dataset (demographic bias) encompasses a varied collection of scenarios, each depicting a hypothetical case where a decision is required (Appendix F.2).

3.1 Gender Bias: WinoBias

The sentences in WinoBias are designed to be structurally parallel but differ in the gender pronouns. It contains two sets of sentences: *pro sentences* with the pro-stereotypical gender pronouns (e.g., nurses as she, engineers as he), and *anti sentences* with anti-stereotypical gender pronouns (e.g., nurses as he, engineers as she). The dataset also has two types of tasks with different levels of difficulties: coreference decisions in Type I task are challenging and must be made using world knowledge about given circumstances, whereas Type II task can be resolved using only syntactic information.

For each sentence in WinoBias, we define the `original question` as "Who does [gender pronoun] refer to in the sentence '[original sentence]'?", and we measure the performance of four large language models across the above two types of coreference tasks (Type I and Type II).

Baselines and Evaluation Metrics There are a few existing works employing prompting techniques to mitigate the bias in LLMs, and we have listed our baselines in Appendix **??**. We measure the performances of LLMs on *pro* and *anti sentences* in terms of accuracy, and the gap between the two indicates the level of gender bias exhibited by the models (a smaller gap is better).

Our Method We propose `Reduce + Fact` as one way of encouraging bias-free reasoning collectively with reducing biased reasoning. As shown in Figure 1(b), we first create two gender-agnostic sentences by replacing the gender pronoun with the two occupations that appeared in the original sentence. Then we ask the model a `factual question`: which sentence is more likely to happen in real life, as this question does not contain any gender-related information. We then include its answer when asking the `original question`. This allows us to distill LLM's non-gender-related world knowledge and nudge it to explicitly reason with this knowledge during coreference resolution (Strategy I). One caveat of using LLM's world knowledge is that the performance of `Reduce + Fact` will increase as the capabilities of the LLM grow, since better world knowledge will further help its performance on both types of sentences. On top of encouraging bias-free reasoning (approach `Fact`), we also tell the model that both occupations are equally likely to be male or female to counteract its existing selection bias (Strategy II for `Reduc[ing]` biased reasoning).

Main Results (Type I Task) In Table 1, the `Default` prompting shows significant biases, with large gaps in accuracy between *pro sentences* and *anti sentences* for all models. The method `Zero-shot COT`, marginally reduces the bias, as seen in the smaller gaps, but its performance is lower on both pro and anti scenarios for GPT-3 and GPT-3.5 models when compared with `Default`, and it only marginally improved the general performance when applied with more capable models (Claude 2 and GPT-4). For `ICL with constrastive examples`, although the gap became larger on less capable models (GPT-3 and GPT-3.5), with more capable LLMs (Claude 2 and GPT-4), it can further reduce bias, especially in GPT-4 with a gap of 9.23%.

Table 1: **Performance comparison of various debiasing methods on WinoBias.** We show that combining *encouraging bias-free reasoning* and *reducing biased reasoning* together (`Reduce + Fact`) greatly decreases the gender bias on the Type I coreference task, which requires world knowledge. *Pro* stands for coreference with pro-stereotypical pronouns, and *Anti* stands for anti-stereotypical pronouns. A lower gap indicates less bias.

Accuracy (%)	GPT 3			GPT 3.5			Claude 2.0			GPT 4		
	Anti	Pro	$\text{Gap}_{\downarrow}$	Anti	Pro	$\text{Gap}_{\downarrow}$	Anti	Pro	$\text{Gap}_{\downarrow}$	Anti	Pro	$\text{Gap}_{\downarrow}$
Default	43.01	79.24	36.23	62.96	94.03	31.07	67.57	92.13	24.56	82.50	97.96	15.47
COT (zero shot)	41.79	75.85	34.06	62.14	90.64	28.49	70.56	91.59	21.03	84.40	95.66	11.26
ICL	46.81	94.57	47.76	45.18	92.81	47.63	73.68	92.27	18.59	88.87	98.10	9.23
Reduce + Fact	73.27	73.95	**0.68**	72.73	84.67	**11.94**	74.08	75.17	**1.09**	94.57	96.74	**2.17**

Remarkably, `Reduce + Fact`, which encourages bias-free reasoning with the reduction of biased reasoning, substantially decreases the bias across all LLMs. This is evidenced by the minimal gaps, with GPT-4 exhibiting a mere 2.17% gap and 94.57% accuracy on *anti sentences*, signifying a significant debiasing effect when compared to other types of prompt designs.

These experimental results suggest prompt designs that encourage bias-free reasoning and (or) reduce biased reasoning are effective at mitigating gender biases in large language models by directing them to rely more on non-gender-related world knowledge and less on gender shortcuts, thus promoting fairer and less biased responses. Also, the performance gap between *pro sentences* and *anti sentences* decreases as the LLMs become more capable, which may indicate that LLMs are less prone to assign occupations with stereotypical gender pronouns as their general (reasoning) capabilities grow.

4 Conclusion

This paper presents a causality-guided and prompting-based LLM debiasing framework. In particular, we highlight the key role of *selection mechanisms* in modeling data corpus bias and in formulating how prompt designs can influence LLM outputs by specifying different selection conditions on its internal representations. Guided by causal understandings of such interplay, we identify principled debiasing prompting strategies. Our strong empirical results demonstrate the benefits of our framework, offering clear intuitions and theoretical foundations for effective debiasing approaches. Future work will naturally extend to acquiring bias-free knowledge and representations for LLMs.

Acknowledgement

This material is based upon work supported by the AI Research Institutes Program funded by the National Science Foundation (NSF) under AI Institute for Societal Decision Making (AI-SDM), Award No. 2229881. This project is also partially supported by an Amazon Research Award (Fall 2022 CFP), the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Apple Inc., KDDI Research Inc., Quris AI, and Infinite Brain Technology.

REFERENCES

Abubakar Abid, Maheen Farooqi, and James Zou. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463, 2021.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pp. 100–108. PMLR, 2012.

Elias Bareinboim and Jin Tian. Recovering causal effects from selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021.

Joseph Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349–4357, 2016.

Shikha Bordia and Samuel R Bowman. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*, 2019.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7801–7808, 2019.

Juan Correa and Elias Bareinboim. General transportability of soft interventions: Completeness results. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10902–10912, 2020.

Juan D Correa, Jin Tian, and Elias Bareinboim. Identification of causal effects in the presence of selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2019.

Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. In *International Conference on Machine Learning*, pp. 2185–2195. PMLR, 2020.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*, 2021.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.

Robert F Engle, David F Hendry, and Jean-Francois Richard. Exogeneity. *Econometrica: Journal of the Econometric Society*, pp. 277–304, 1983.

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

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, N Anandhavelu, Niyati Chhaya, and Balaji Vasan Srinivasan. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4534–4545, 2021.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.

Yue Guo, Yi Yang, and Ahmed Abbasi. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1012–1023, 2022.

James Heckman. Varieties of selection bias. *The American Economic Review*, 80(2):313–318, 1990.

James J Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, pp. 153–161, 1979.

Miguel A Hernán and James M Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(1):3482–3534, 2020.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*, 2023.

David Kaltenpoth and Jilles Vreeken. Identifying selection bias from observational data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Masahiro Kaneko and Danushka Bollegala. Debiasing pre-trained contextualised embeddings. *arXiv preprint arXiv:2101.09523*, 2021.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 656–666, 2017.

Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. Sustainable modular debiasing of language models. *arXiv preprint arXiv:2109.03646*, 2021.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 1931–1940, 2018.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In *International Conference on Machine Learning*, pp. 4674–4682. PMLR, 2019.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Optimal training of fair predictive models. In *Conference on Causal Learning and Reasoning*, pp. 594–617. PMLR, 2022.

Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. In-contextual bias suppression for large language models. *arXiv preprint arXiv:2309.07251*, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 2023.

Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30 (4):962–1030, 2002.

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

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

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*, 2019.

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

Peter Spirtes. Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 491–498, 1995.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer New York, 1993.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in Artificial Intelligence*, pp. 499–506, 1995.

Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10–29, 2013.

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

Zeyu Tang, Yatong Chen, Yang Liu, and Kun Zhang. Tier balancing: Towards dynamic fairness over underlying causal factors. In *International Conference on Learning Representations*, 2023a.

Zeyu Tang, Jiji Zhang, and Kun Zhang. What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. *ACM Computing Surveys*, 55(13s):1–37, 2023b.

Zeyu Tang, Jialu Wang, Yang Liu, Peter Spirtes, and Kun Zhang. Procedural fairness through decoupling objectionable data generating components. In *International Conference on Learning Representations*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9584–9594, 2022.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*, 2023a.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Christopher Winship and Robert D Mare. Models for sample selection bias. *Annual Review of Sociology*, 18(1):327–350, 1992.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *arXiv preprint arXiv:2310.10683*, 2023.

Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, et al. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint arXiv:2304.05524*, 2023.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

Kun Zhang, Jiji Zhang, Biwei Huang, Bernhard Schölkopf, and Clark Glymour. On the identifiability and estimation of functional causal models in the presence of outcome-dependent selection. In *UAI*, 2016.

Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? In *Advances in Neural Information Processing Systems*, volume 33, pp. 18457–18469, 2020.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2, 2018.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*, 2023.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 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)*, pp. 4227–4241, 2023.

TABLE OF CONTENTS: APPENDIX

A Related Works

In this section, we provide detailed discussions on related works. We consider the combinations of very related topics including causality, algorithmic fairness, and LLM reasoning. In Section A.1, we consider causal notions of fairness that do not specifically pertain to the LLM context. In Section A.2, we consider existing efforts to draw connections between causality and LLM reasoning. In Section A.3, we consider previous works on LLM debiasing. In Section A.4, we consider previous works that involve all three topics.

A.1 Causality and Fairness

It has been recognized in the algorithmic fairness literature that causality provides a unique tool to facilitate a better understanding of the data-generating process, and therefore, more effective bias quantifications and mitigations (Kilbertus et al., 2017; Kusner et al., 2017; Nabi & Shpitser, 2018; Nabi et al., 2019; Chiappa, 2019; Nabi et al., 2022; von Kügelgen et al., 2022; Tang et al., 2023a). Previous causal fairness literature has considered notions based on estimating or bounding various kinds of causal effects (Kilbertus et al., 2017; Kusner et al., 2017; Nabi & Shpitser, 2018; Nabi et al., 2019; Chiappa, 2019), and also the causal modeling of the dynamics as well as the long-term implications of bias mitigation strategies (Creager et al., 2020; Zhang et al., 2020; Tang et al., 2023a;b).

Because of the relatively abstract definition of the variable or node in the language context, previous approaches for characterizing and enforcing causal fairness are not directly applicable in LLM debiasing tasks. That being said, as we have demonstrated in our causality-guided debiasing framework, causal understandings of the involved data-generating processes help identify effective debiasing strategies that are both intuitively clear and theoretically grounded.

A.2 Causality and LLMs

The intersection between causality and LLMs has drawn increasing attention. Zhang et al. (2023) considers three types of causal questions and aims to evaluate LLMs' abilities to identify causal relations, discover new knowledge from data, and quantitatively estimate the consequences of actions. Kıcıman et al. (2023) investigate LLMs' abilities to perform causal reasoning and solve covariance-/logic- based causal questions. They also study the failure modes of LLMs and provide techniques to interpret the model robustness. Jin et al. (2023) propose a benchmark data set for evaluating LLMs' causal inference capabilities via the task of determining causal relationships from a set of correlational statements.

This line of research focuses on complex causal reasoning abilities in general settings, without specific attention to potential fairness violations. In comparison, our causality-guided debiasing framework does not involve assumptions/requirements on LLMs' general-purpose causal reasoning capabilities. We adopt a rather mild assumption that a well-trained and well-aligned LLM captures the dependence pattern in the training data and that such a pattern is internalized and utilized during reasoning.

A.3 Debiasing Language Models

There is a large amount of work discussing bias and fairness in the context of language models (LMs) Bordia & Bowman (2019); **?**); Abid et al. (2021); Wang et al. (2023a); Liu et al. (2023); Ray (2023); Rozado (2023), and our investigation lies on debiasing techniques with causal understandings of the sources of biases. For debiasing approaches in the context of LMs, there are proposals involving direct fine-tuning of model parameters (Kaneko & Bollegala, 2021; Garimella et al., 2021; Lauscher et al., 2021; Guo et al., 2022), modifying the decoding steps (Schick et al., 2021), incorporating Reinforcement Learning with Human Feedback (RLHF) to better align the models with human values (Ouyang et al., 2022; Bai et al., 2022; Yao et al., 2023), and prompting-based techniques (Si et al., 2022; Tamkin et al., 2023; Oba et al., 2023; Ganguli et al., 2023). We focus on prompting-based techniques and identify principles for prompt designs to steer LLMs toward unbiased responses by a). reducing biased reasoning and b). encouraging bias-free reasoning. We provide demonstrations of how we can employ the above two principles. Works on LLMs reasoning such as Wei et al.

(2022); Zheng et al. (2023) can also be incorporated into with our framework to encourage bias-free reasoning.

A.4 Debiasing LLMs from Causal Perspectives

Most closely related literature considers LLM debiasing strategies from causal perspectives. Vig et al. (2020) utilize causal mediation analysis and consider neuron-level intervention to investigate the instantiation of gender bias in Transformer-based language models (Vaswani et al., 2017). Zhou et al. (2023) propose `Causal-Debias` to mitigate the unwanted stereotypical association by fine-tuning pretrained language models. The causal model they consider involves four variables: label-relevant factor, bias-relevant factor, raw sentence, and ground-truth label. Wang et al. (2023b) pay special attention to entity bias and propose a specific structural causal model (SCM) for easier parameter estimations, such that the intervention-based mitigation strategy can be carried out. Their causal model involves four variables: entity, raw text, LLM input, and LLM output.

In comparison, we provide detailed causal modelings at a sub-sentence level, considering both the training corpus generating process and the LLM reasoning process. Furthermore, our framework explicitly models the interplay between internal representations and external inputs through selection mechanisms, providing a clearer picture regarding possible strategies to debias LLM outputs more effectively. Our approach does not require white-box access or the ability to perform interventions, making our prompting-based framework applicable to a variety of practical scenarios.

B Preliminaries

In this section, we provide a brief introduction to causal modeling and causal reasoning (Section B.1). We also provide a motivating example to illustrate how the selection mechanism can reshape dependence patterns within the different data (Section B.2).[3]

B.1 A Brief Introduction to Causality

For two random variables X and Y, X is a cause of Y if there is a change in the distribution of Y when we apply an intervention on X while holding all other variables fixed (Spirtes et al., 1993; Pearl, 2009). We can represent causal relations among variables with a directed acyclic graph (DAG), where nodes represent variables, and edges represent direct causal relations between variables. We denote the direct causal relation between the ordered pair (X, Y) by a directed edge $X \to Y$.

Local causal modules in a DAG, which characterize the causal relations between the corresponding variable and its direct causes, do not interfere with each other because of causal modularity. This property is also known as the exogeneity (Engle et al., 1983), or the independence of causal mechanism (Peters et al., 2017), resulting directly from the causal Markov condition for the DAG (Spirtes et al., 1993; Pearl, 2009).[4] In the context of language processing, the definition of a variable representing text or tokens is relatively abstract compared to the statistical notion of a random variable in tabular data. Within the scope of this work, we use the terms "variable" and "node" interchangeably when the context permits clear understandings.

B.2 A Motivating Example of Selection Mechanisms

Ideally, we would like samples to be drawn uniformly from the underlying population of interest. However, in practice, it is very common that the probability of including certain data points in the corpus depends on the characteristics of the data points themselves.

Previous literature has investigated selection mechanisms from different perspectives, for instance, the influence of selection bias on statistical inference in economic and sociological studies (Heckman, 1979; 1990; Winship & Mare, 1992), causal discovery when there are selection variables and latent

[3]We discuss related works in detail in Appendix A.

[4]There are additional classes of graphs considered in causality literature, for example, directed cyclic graphs (DCGs) (Spirtes, 1995), ancestral graphs (Richardson & Spirtes, 2002), and so on. In this paper, we consider causal processes that can be modeled by a DAG. Other graph classes are beyond the scope of our work.

common causes (Spirtes et al., 1995; Zhang, 2008), the identification and estimation of functional causal models when selection exists (Zhang et al., 2016), the identifiability of causal effect in the presence of selection bias from the graphical condition perspective (Bareinboim & Pearl, 2012; Bareinboim & Tian, 2015; Correa et al., 2019) and from the potential outcome perspective (Hernán & Robins, 2020), and the identification of the existence of selection bias from observational data under certain functional assumptions (Kaltenpoth & Vreeken, 2023).

Let us consider an example in the context of medical care. As illustrated in Figure 3, observed variables (X_1, X_2) denote diseases, and (Y_1, Y_2) denote corresponding symptoms. Apart from them, there exist potential binary selection variables S_i's ($i \in \{1, 2, \ldots, 5\}$), where $S_i = 1$ denotes being selected. X_1 (X_2) is the direct and only cause of Y_1 (Y_2), and the two diseases are unrelated in the general population, i.e., when none of the S_i's exists. We use solid edges to represent causal relations among observed variables and dashed edges for those pertaining to selection mechanisms.

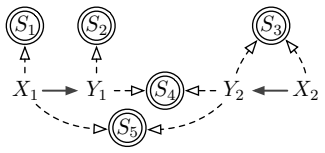


Figure 3: An illustrative example of selection mechanisms.

For instance, S_2 is an example of outcome-dependent selection (Zhang et al., 2016), which can be selecting individuals with symptom Y_1 from the general population. The conditional probability $P(Y_1 \mid X_1, S_2 = 1)$ in the selected data typically differs from its counterpart $P(Y_1 \mid X_1)$ in general population (Zhang et al., 2016). As another example, S_4 can denote the selection mechanism of only considering the hospital in-patient data. This signifies the setting of the well-known Berkson's Paradox (Berkson, 1946), where two unrelated diseases (X_1, X_2) appear to be correlated in the hospital data, simply because the data only contains selected patients who have at least one symptoms in (Y_1, Y_2).

Figure 3 shows that the selection can be based solely on the cause (e.g., S_1), solely on the effect (e.g., S_2), or on both (e.g., S_3). Meanwhile, the selection can be based on variables that are not causally related in the general population (e.g., S_4 and S_5). Selection mechanisms can reshape dependence patterns among involved variables, and as a consequence, potentially change downstream outputs as well. We will see in Section 2 that such property of selection also applies to natural language processing (NLP) contexts.

C Debiasing Guided by Causal Understandings

C.1 Remark on the Modeled Processes

Figure 2(a) and Figure 2(b) each have their own generating process of interest and distinct emphases. Although not pertaining to LLM reasoning itself, the underlying generating process modeled in Figure 2(a) provides hints on local causal modules of interest that prompt designs can specifically attend to for debiasing purposes. When we consider these hints with reference to the LLM reasoning process modeled in Figure 2(b), we can identify debiasing strategies that are both intuitive and theoretically grounded, as we will see in more detail in Section C. The two detailed causal models complement each other, both of which are essential for understanding the source of bias, and furthermore, effectively debiasing LLMs.

C.2 Prompt-based LLM Debiasing Frameworks

We present our prompt-based LLM debiasing framework guided by causal understandings of the related data generating processes. Our core idea is to formulate conditions that should be specified in the prompt design, such that through the influence of selection mechanisms on LLM reasoning process, one can effectively debias LLM outputs.

Figure 2(c) presents an annotated version of the LLM reasoning process (Section 2). Based on the understanding of the underlying generating process of training data corpus, and the mild assumption that the trained LLM captures the dependence patterns in training data, we highlight certain edges in light coral (accompanied by annotations marked with circled red numerals) in Figure 2(c) to denote unregulated information flow from demographic representations to LLM outputs in the internal reasoning process. We use circled blue Roman numerals to denote selection mechanisms that can

be specified by the external input prompt (Section 2), and circled red Roman numerals to represent those corresponding to historical discriminations (Section 2).

According to LLM reasoning process presented in Figure 2(c), we consider the information flows from the demographic representation (upstream node) to the LLM potential output (downstream node). We present additional conditions and constraints the prompt designs should specify for debiasing purposes. To clearly present the intuitions and theoretical groundings of our causality-guided framework, we identify three prompting strategies for debiasing LLMs, each of which serves as a solid starting point.

Strategy I (*Nudge Towards Demographic-Agnostic Fact*). The intuition behind is to nudge LLMs towards utilizing demographic-agnostic fact when generating the output.

Condition I The internal representations for demographic-agnostic fact and demographic information should be conditionally independent in the presence of PPC and existing selection S:

$$\begin{matrix}\text{demographic}\\ \text{representation}\end{matrix} \perp\!\!\!\perp \begin{matrix}\text{demographic-agnostic}\\ \text{fact representation}\end{matrix} \mid S = 1, \text{PPC} = 1. \tag{1}$$

Strategy I introduces Condition I to specify the selection mechanism over internal representations of demographic information and demographic-agnostic fact, denoted by edges (i) and (v) in Figure 2(c).

Strategy II (*Counteract Existing Selection Bias*). The intuition behind this strategy is to directly counteract the effect of existing historical discriminations.

Condition II.1 The internal representations for demographic information and entity should be conditionally independent in the presence of PPC and existing selection S:

$$\begin{matrix}\text{demographic}\\ \text{representation}\end{matrix} \perp\!\!\!\perp \begin{matrix}\text{entity}\\ \text{representation}\end{matrix} \mid S = 1, \text{PPC} = 1. \tag{2}$$

Condition II.2 No new association between internal representations of demographic information and demographic-agnostic fact is introduced:

$$\begin{matrix}\text{demographic}\\ \text{representation}\end{matrix} \perp\!\!\!\perp \begin{matrix}\text{demographic-agnostic}\\ \text{fact representation}\end{matrix} \mid \begin{matrix}\text{entity}\\ \text{repr.}\end{matrix}, \begin{matrix}\text{scenario}\\ \text{repr.}\end{matrix}, S = 1, \text{PPC} = 1. \tag{3}$$

In Strategy II, Condition II.1 and Condition II.2 serve different purposes. Condition II.1 aims to counteract existing bias instantiated by selection S (edges (vi) and (vii)) by constraining marginal dependence between representations for demographic information and entity (edges (i) and (ii)). Condition II.2 acts as a safeguard, making sure no new bias is introduced in the causal downstream of the entity so that the above counteraction effectively proceeds to the final output.

Strategy III (*Nudge Away from Demographic-Aware Text*). The intuition behind this strategy is to nudge LLMs away from utilizing demographic-aware text to generate outputs.

Condition III The internal representations for demographic-aware text and demographic information should be conditionally independent in the presence of PPC and existing selection S:

$$\begin{matrix}\text{demographic}\\ \text{representation}\end{matrix} \perp\!\!\!\perp \begin{matrix}\text{demographic-aware}\\ \text{text representation}\end{matrix} \mid S = 1, \text{PPC} = 1. \tag{4}$$

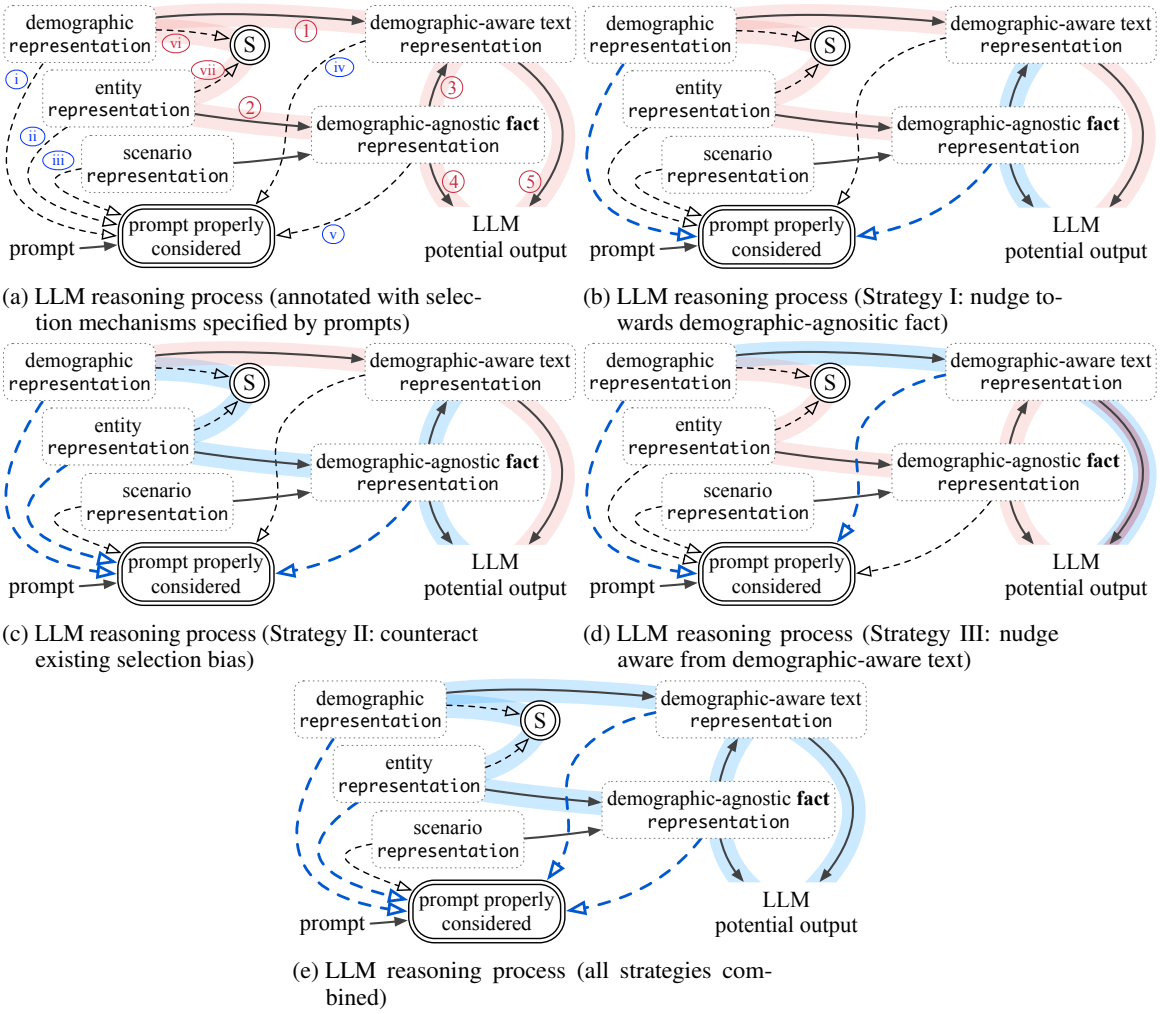
Strategy III utilizes Condition III to specify the selection mechanism over internal representations of demographic information and demographic-aware text (edges (i) and (iv)) to regulate the information flow along the edge (1).

Remarks on Three Strategies

The three strategies offer certain effectiveness individually but are not perfect on their own.[5] While Strategy I nudges LLMs towards utilizing demographic-agnostic facts, it does *not* explicitly prevent LLMs from using demographic-aware text representations to generate the output, and the demographic information can potentially be associated to the output through an unregulated path containing edges (1) and (5). Similarly, while Strategy II aims to regulate information flows along edges {(2), (3), (4), (vi), (vii)}, there is *no* explicit constraint involving edges (1) and (5), which leaves space for bias to sneak in during the reasoning process.

Compared to Strategy I that pushes LLMs to focus on demographic-agnostic facts, Strategy III prevents LLMs from referring to demographic-aware text during reasoning. While edge (1) is explicitly regulated by the selection mechanism specified by Condition III, the information flow

[5]We provide additional illustrations in Appendix D.



(a) LLM reasoning process (annotated with selection mechanisms specified by prompts)

(b) LLM reasoning process (Strategy I: nudge towards demographic-agnositic fact)

(c) LLM reasoning process (Strategy II: counteract existing selection bias)

(d) LLM reasoning process (Strategy III: nudge aware from demographic-aware text)

(e) LLM reasoning process (all strategies combined)

Figure 4: Additional illustrations on debiasing strategies.

along edge (5) can still result in the association between demographic information and the output. This is because Strategy III does *not* involve conditions or constraints over the dependence patterns among variables including demographic information and demographic-agnostic fact. As we can see from the conditions specified in Strategy II, no explicit appearance of demographic information in a demographic-agnostic fact does not guarantee the non-existence of dependence between the two. We would like to note that debiasing is better realized when the strategies are combined as they can address social bias in LLMs more comprehensively.

D Further Illustrations and Discussions of Our Framework

In this section, we provide further illustrations on the debiasing strategies identified in Section C.

In Figure 4, we use light coral (blue) to highlight unregulated (regulated) information flows from demographic representations to LLM outputs in the internal reasoning process. We use mix-colored highlight to denote partially regulated information flow, e.g., the edge from demographic-aware text representation to LLM potential output in Figure 4(d). Comparing Figure 4(e) (all strategies combined) with Figures 4(b) – 4(d) (strategies applied individually), we can see that collectively, the strategies address the social bias in language more comprehensively.

Here by "regulated", we are referring to the constraint over the association between demographic information representation and the potential output. We would like to note that our causal models can

naturally handle situations where certain kinds of neutral dependence patterns involving demographic information are not necessarily considered as problematic (Tang et al., 2024). We present relevant experimental results in Section F.1. We would also like to note that our causality-guided framework is not limited to situations where demographic information is explicitly involved. One can adapt our framework to the specific practical setting by incorporating relevant nodes into causal graphs, thereby identifying the most suitable debiasing strategies therein.

E EXPERIMENTAL DETAILS

Pretrained LLMs For the GPT models used in our experiments, we consider snapshots from June 13th, 2023 where the knowledge cut-off time is Sep 2021. Since the legacy GPT-3 model (a.k.a., text-davinci-003) is no longer supported when we conduct the experiments, we use the model "gpt-3.5-turbo-instruct" instead as it has similar capabilities as GPT-3 era models. The Mistral-7B model we use in our experiments is the improved instruction fine-tuned version (a.k.a., "Mistral-7B-Instruct-v0.2"). We are unable to use Claude 2 for our experiments on the Discrim-Eval dataset because we cannot access the probabilities of generated tokens. All LLMs responses are obtained with a temperature of 0.

Iterative Prompting The responses of LLMs may not follow a given format even with specific instructions, which makes answer extraction challenging. Moreover, we observe that adding instructions to restrict the format of responses may lower the quality of the generated outputs. Therefore, we apply 2-round iterative prompting in our experiments where we let the models generate freely and then ask them to summarize their answers in one or two words. This way allows us to obtain high answer qualities with an easy extraction process at the same time.

E.1 GENDER BIAS: WINOBIAS

For experiments on the WinoBias dataset, we combined both the training and test data for evaluation as there is no need to separate them when using prompting-based debiasing techniques. We also noticed that the original dataset contains 400 grammar mistakes across 3000+ sentences in total. We corrected these grammar mistakes as they may hinder the co-reference resolution tasks. We have provided the cleaned version in the supplementary materials.

Since there are 571 examples where two or more models make the same mistakes on the `factual questions`. After looking into these mistakes, we find that some of the examples are indeed ambiguous in nature (e.g., "The construction worker stared at the teacher because [`who`] was angry"). Therefore, we collected human evaluations to understand the inherent ambiguity in the WinoBias dataset. For each of the 571 examples, we collect 5 human responses on the `factual questions` (i.e., which sentence is more likely to happen in real life), and we ask the human annotator to choose from $\{$Sentence A, Sentence B, and Equally likely$\}$. Based on the human responses, we identified 60 examples (55 from the Type I task and 5 from the Type II task) where three or more annotators disagree with the ground truth answers or think both sentences in the `factual questions` are equally likely to happen in real life. We removed these 60 examples during our evaluation, and we will include the human evaluation results in our codebase.

E.2 DEMOGRAPHIC BIAS: DISCRIM-EVAL

The Discrim-Eval dataset contains 70 diverse decision scenarios and $9 \times 3 \times 5 \times 70 = 9450$ individual decision questions which includes all combinations of [AGE] $\in$ [20, 30, 40, 50, 60, 70, 80, 90, 100], [GENDER] $\in$ [male, female, non-binary] and [RACE] $\in$ [white, Black, Asian, Hispanic, Native American]. To measure the corresponding bias in each demographic category, we reconstruct the dataset by extracting the `base scenario` which does not contain any demographic information (e.g., we replace all pronouns with the anaphoric reference to avoid leaking the gender information). We then ask the model to decide on each of the 70 `base scenarios`. There are (1/11/1/2) scenarios where (Mistral 7B/GPT-3/GPT-3.5/GPT-4) refuses to answer or does not output a Yes answer, and we removed these scenarios correspondingly when evaluating these LLMs.

Table 2: **Error Analysis on WinoBias Type II coreference task.** We divide the models' responses into 4 categories to better understand their success and failure cases. **TT** denotes the (%) of examples where LLM answers both the factual question and the original question correctly, **FF** denotes the (%) of examples where both questions are answered incorrectly, indicating the coreference errors caused by the model's **world knowledge**. **TF** denotes the coreference errors caused by **gender bias** as only factual questions are correctly answered, and **FT** indicates coreference success that may be due to **gender shortcut** since the factual questions get wrong but original questions are correctly answered.

Accuracy (%)	TT		TF		FT		FF	
	Anti	Pro	Anti	Pro	Anti	Pro	Anti	Pro
GPT-3.5								
Reduce + Fact	88.69	89.71	1.14	0.13	3.43	4.70	6.73	5.46
Fact Only	87.55	89.83	2.29	0.00	3.05	5.97	7.12	4.19
Reduce Only	80.30	85.77	5.72	2.03	8.77	9.15	0.89	0.51
Default	83.61	89.71	5.59	0.13	8.64	10.04	1.40	0.13
GPT-4								
Reduce + Fact	99.62	99.62	0.13	0.00	0.00	0.13	0.25	0.25
Fact Only	99.36	99.87	0.25	0.00	0.13	0.00	0.25	0.13
Reduce Only	98.98	99.49	0.76	0.13	0.00	0.38	0.25	0.00
Default	98.73	99.75	0.89	0.13	0.25	0.00	0.13	0.13

Table 3: **Error analysis of Reduce + Fact on Type I and Type II questions in WinoBias.**

Accuracy (%)	TT		TF		FT		FF	
	Anti	Pro	Anti	Pro	Anti	Pro	Anti	Pro
GPT-3								
Type I	72.73	73.27	1.49	0.95	0.54	0.68	25.24	25.10
Type II	82.47	82.34	1.27	1.40	0.76	1.91	15.50	14.36
GPT-3.5								
Type I	70.15	78.70	10.58	2.04	2.58	5.97	16.55	13.16
Type II	88.69	89.71	1.14	0.13	3.43	4.70	6.73	5.46
Claude								
Type I	73.00	73.00	2.99	2.99	1.09	2.17	22.93	21.85
Type II	79.80	81.19	6.61	5.21	2.54	4.07	11.05	9.53
GPT-4								
Type I	94.44	96.07	1.76	0.27	0.14	0.68	3.66	2.99
Type II	99.62	99.62	0.13	0.00	0.00	0.13	0.25	0.25

F Additional Experimental Results

F.1 Gender Bias: WinoBias

We include additional ablation studies on the WinoBias dataset, which include a detailed error analysis on WinoBias Type II coreference task (Table 2), a detailed error analysis of Reduce + Fact with 4 LLMs on Type I and Type II tasks (Table 3), and an ablation study on adjusting the levels of biased reasoning (Table 4) in the prompt design.

In Table 3, we show the detailed results on both Type I and Type II coreference tasks across 4 LLMs. As we can see, our method has bigger improvements on models with better world knowledge as. models with worse world knowledge could limit our method to reaching its full capacity.

Ablation study on adjusting the levels of counteracting existing selection bias In Table 4, we investigated how different levels of enforcing Strategy II (counteract existing selection bias) impacts the model debiasing performance by adjusting the Fact part in the input prompt. For example, a counteract level of 100% indicates the following prompt: `Assume that` **`the physician`** `can be male 0% of the time and female 100% of the time, and assume that` **`the secretary`** `can be male 0% of the time and female 100% of the`

Table 4: **Adjusting the levels of counteracting existing selection bias on GPT-3.5 with Reduce + Fact.** We investigated how the level of counteracting stereotype impacts the model performance by adjusting the Fact part in the input prompt. For example, a counteract level of 100% indicates the following prompt: `Assume that` **`the physician`** `can be male 0% of the time and female 100% of the time, and assume that` **`the secretary`** `can be male 0% of the time and female 100% of the time.`

Counteract level (%)	TT		TF		FT		FF	
	Anti	Pro	Anti	Pro	Anti	Pro	Anti	Pro
100	70.15	77.20	10.58	3.66	3.80	4.34	15.47	14.79
90	69.06	77.88	11.53	2.99	1.76	4.61	17.64	14.52
75	68.79	78.02	11.94	2.85	2.17	5.02	17.10	14.11
50	65.94	78.02	15.20	2.99	2.44	5.02	16.42	13.98
25	65.67	80.05	15.20	0.95	1.22	5.43	17.91	13.57
10	64.45	79.78	16.01	1.09	0.81	5.43	18.72	13.70
0	61.06	78.97	19.95	1.90	1.09	8.28	17.91	10.85

`time`; while a level of 50% indicates the following prompt: `Assume that` **`the physician`** `can be male 50% of the time and female 50% of the time, and assume that` **`the secretary`** `can be male 50% of the time and female 50% of the time`. We observe that as the level of anti-stereotype goes down, the errors caused by the use of the gender shortcut increase (**TF** increases). In addition, by soft adjustment of reducing biased reasoning, we provide not only flexible tuning strategies for the best model performance but also a chance to dive into the underlying reasons for the error.

F.2 Demographic Bias: Discrim-Eval

In Figure 5, Figure 6, Figure 7, Figure 8, we show in details the performance comparison on Discrim-Eval across three demographic categories across four LLMs (Mistral 7B, GPT 3, GPT 3.5, and GPT 4). The height of the bar denotes the degree of discrimination by comparing the least privileged group with the most privileged group in a given demographic category (the higher the bar, the deeper the discrimination). Different methods (prompt designs) are colored differently (lighter colors denote the ones that amplify bias-free reasoning). The baseline setting is colored black.

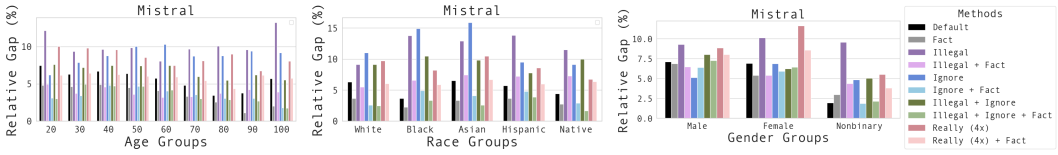


Figure 5: **Performance comparison on Discrim-Eval across three demographic categories in details on Mistral (7B).** The height of the bar denotes the degree of discrimination by comparing the least privileged group with the most privileged group in a given demographic category (the higher the bar, the deeper the discrimination). Different methods (prompt designs) are colored differently (lighter colors denote the ones that amplify bias-free reasoning). Amplifying bias-free reasoning universally reduces the relative gap when added with methods that reduce biased reasoning.

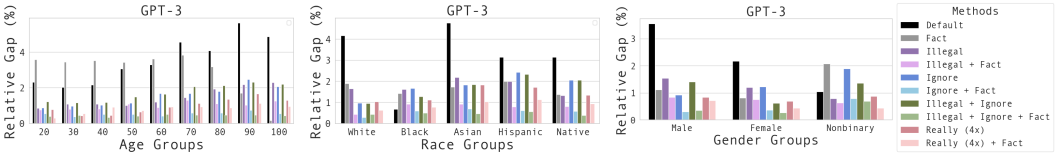


Figure 6: **Performance comparison on Discrim-Eval across three demographic categories in details on GPT-3.** The height of the bar denotes the degree of discrimination by comparing the least privileged group with the most privileged group in a given demographic category (the higher the bar, the deeper the discrimination). Different methods (prompt designs) are colored differently (lighter colors denote the ones that amplify bias-free reasoning). Amplifying bias-free reasoning universally reduces the relative gap when added with methods that reduce biased reasoning.

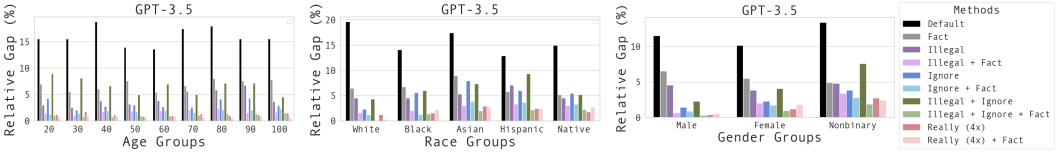


Figure 7: **Performance comparison on Discrim-Eval across three demographic categories in details on GPT-3.5.** The height of the bar denotes the degree of discrimination by comparing the least privileged group with the most privileged group in a given demographic category (the higher the bar, the deeper the discrimination). Different methods (prompt designs) are colored differently (lighter colors denote the ones that amplify bias-free reasoning). Amplifying bias-free reasoning universally reduces the relative gap when added with methods that reduce biased reasoning.

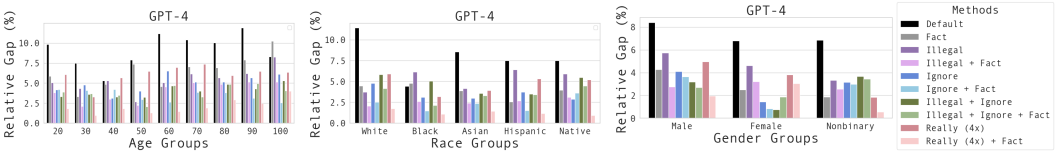


Figure 8: **Performance comparison on Discrim-Eval across three demographic categories in details on GPT-4.** The height of the bar denotes the degree of discrimination by comparing the least privileged group with the most privileged group in a given demographic category (the higher the bar, the deeper the discrimination). Different methods (prompt designs) are colored differently (lighter colors denote the ones that amplify bias-free reasoning). Amplifying bias-free reasoning universally reduces the relative gap when added with methods that reduce biased reasoning.