An Overview of Gender and Political Biases in LLMs

Jared B. Jones
Department of Computer Science
Baylor University
Waco, TX 76710
jared_jones2@baylor.edu

Naeem Seliya, Emily M. Hastings,
Benjamin T. Fine
Computer Science Department
University of Wisconsin-Eau Claire
Eau Claire, WI 54701
seliyana@uwec.edu
hastinem@uwec.edu
finebt@uwec.edu

Abstract

As the public increasingly relies on Large Language Models (LLMs) like GPT for information, and as students increasingly turn to these models for assistance in studying, the impact of their influence will rise. Thus, it becomes increasingly imperative to study what biases these models exhibit. We examine eleven studies to determine the political leanings and the prevalence of gender stereotyping in GPT. Examinations reveal a progressive political bias with varying consistency based on language implementation. Current literature suggests GPT also exhibits gender biases, especially in the choice of adjectives, professions, and gender-based preferences. GPT often favors stereotypical roles or expressions and shows a tendency to assign higher scores or use more positive language when evaluating certain genders. It is likely all the aforementioned biases reflect the biases inherent in the training data. To mitigate these biases, we suggest training LLMs with diverse training data incorporating a balance of contrasting perspectives.

1 Introduction

Bias is an inclination that affects the accuracy, credibility, and objectivity of information sources (Burkholder et al., 2022). One area where this inclination is particularly of note is the topic of politics. Political information often varies significantly based on the biases of its source leading its consumers with contrasting perspectives. As GPT grows as a source for political information both historical and current, its potential to exhibit its biases on a large population of individuals increases. This influence may result in the political perspective of users being shaped in part by the chatbot, resulting in a shift in perspective toward GPT's political leanings (Rotaru et al., 2024). This shaping could lead students to develop a narrow and polarized understanding of the world while encouraging intolerance and close-mindedness (Powers et al., 2019).

Gender bias exhibited in a source of guidance like GPT could also result in the perpetuation of stereotypes and attitudes that marginalize, and disadvantage individuals based on their gender (Ellemers et al., 2018). The impact of bias is especially important to consider for groups of the population such as students. For this demographic, bias could influence self-esteem and career aspirations while confining students to narrow expectations and reducing equality (Frawley, 2005).

To help provide further context for this issue, we conducted a literature review seeking to answer the following questions: (1) What are the political leanings of GPT? and (2) To what extent is gender stereotyping prevalent in GPT? We review literature to examine studies conducting political orientation tests on GPT to understand its political biases. In addition to political bias, we analyze how GPT might exhibit gender bias by varying word choice based on gender, such as assigning traits or characteristics that are consistent with stereotypes. We looked at how different linguistic versions of GPT perpetuate both gender and political bias.

We conclude that GPT tends to assign phrases, words, and professions in line with gender stereotypes. GPT also appears to have a consistent left-wing political bias. There is also evidence to suggest that GPT varies its political claims and objectivity based on the language version of GPT (Kuznetsova et al., 2023; Zhou & Zhang, 2023). These biases likely originate from the biases in GPT's training data (Agiza et al., 2024). To mitigate biases, we suggest training LLMs on more balanced and representative data.

2 Related Work

Several types of biases have been examined in large language models (LLMs), including political, gender, racial, national, linguistic, religious and more. Hu & Rangwala (2020) find bias against male and African American students in machine learning models used in educational settings for tasks such as predicting at-risk students. Similarly, Zhang et al. (2023) prompted GPT for music and movie recommendations without providing sensitive information and contrasted with recommendations that include sensitive information like gender and race. The study showed that GPT was unfair by varying recommendations based on this information. Additional studies have found evidence of gender bias in LLMs generally (Seo et al, 2022, Zhang & Zhou, 2024). There is also evidence to suggest the presence of linguistic bias: LLMs provided lower scores for German speakers than Japanese speakers compared to scores given by human evaluators using a rubric covering aspects of language proficiency, including delivery, language use, and content (Ohi et al., 2024; Loukina et al., 2019). GPT may also vary positive associations based on race and religion (Nadeem et al., 2020). There is further evidence that GPT may predict an individual's nationality based on positive or negative traits and adjust recommendations depending on nationality (Zhang et al., 2023; Kamruzzaman et al., 2023).

Large language models reflect the content of their training datasets (Agiza et al., 2024). These datasets typically represent the most prominent narratives, regardless of whether less represented counternarratives are factually correct (Weidinger et al., 2021). In addition to narratives, datasets may overrepresent certain demographics, perspectives, or ideologies while underrepresenting or excluding others (Bender et al., 2021). Research suggests that the alteration of biases in an LLM's training data will be reflected in the LLM. Agiza et al. (2024) curated LLM training data to deliberately produce left-leaning and right-leaning chatbots and measured political orientation using the Political Compass Test. Han et al. (2024) produced altered datasets for GPT-2 to reduce responses that exhibit stereotypes. The authors used the StereoSet intrasentence dataset (Nadeem et al., 2024) to measure the presence of responses that exhibited stereotypes (SS) and language modeling score (LMS) (Nadeem et al., 2024) to measure the presence of meaningful associations in responses. One of the datasets produced for GPT-2 lowered SS and only slightly lowered LMS (Han et al., 2024).

In this review, we extend this body of prior work by identifying trends related to gender and political bias in GPT and providing potential measures to minimize their impacts.

3 Scope of our Review Study

In accordance with GPT-2's recent release in 2019, this literature overview covers research published from 2020 through 2024. We selected research that prompted GPT with questions from political tests. Secondly, we selected papers that studied the accuracy of GPT political claims. This included both papers that studied the topic generally and ones that held a special emphasis on low-resource language versions of GPT. This paper also covers the presence of gender stereotyping in GPT; specifically, we focused on research works that studied the prevalence of gender stereotypes in word choice. We excluded studies that focused on bias in GPT in other areas besides political and gender bias. We also excluded studies that focused on other LLMs besides GPT.

Ref.	Bias Found	Methodology	Measurement
[6]	Gender and cultural bias in interview evaluations	Using LLMs to grade interviews with variable gender and culture	Comparative analysis of grades across cultural and gender groups
[8]	Gender-occupation biases	Word Embedding Association Test	Analyzing translation outputs for pronoun usage
[11]	Attribute favoritism based on gender	SAI and ASA fill-in-the-blank sentences	Conditional likelihood changes in LLM responses
[12]	Differing of veracity based on language version of GPT	Response examination of socio-political statements	Accuracy and consistency in determining statement veracity
[16]	Biases consistent with stereotypes	Developed the Context Association Test (CAT) to measure biases.	CAT score to quantify how close a model is to an ideal unbiased language mode
[20]	Left-leaning bias in political orientation	Administration of political orientation tests to GPT	Scores from political orientation tests
[22]	Left-leaning and libertarian bias in political orientation	Administration of the political compass test and questionnaires for G7 member states	Scores from political orientation tests
[24]	Variation in positive characteristics based on differing pronouns	Probed ChatGPT with open-ended prompts in English and German	Analysis of word usage in GPT generated responses
[26]	Presence of gender stereotypical language and style in GPT responses	Context-Less Generation and Context-Based Generation	Odds Ratio for word frequency. T-tests for stylistic differences
[27]	Bias toward political viewpoints differing by language version of GPT.	Examination of GPT generated responses to political and natural science questions	Contrasted Chinese and English GPT responses of political and natural science questions
[28]	Differing recommendations based on gender	Examination of GPT generated recommendations for movies and music	Utilized Sensitive-to-Neutral Similarity Range and Sensitive-to-Neutral Similarity Variance to measure unfairness

Table 1: Summary of Literature Findings

4 Review of Select Literature (Gender Bias)

Here we describe the findings of our literature review. See Table 1 for a summary of these findings. In this section, we review the perpetuation of gender stereotypes in GPT. Nadeem et al. (2020) produced a model for large language models like GPT2 to contrast bias with the ability to make meaningful associations. The authors noted a strong correlation between a language's ability to make meaningful associations and the prevalence of bias, suggesting that an LLM's proficiency increases its tendency to exhibit stereotypes. However, they asserted that GPT-2 achieved a favorable balance, exhibiting low bias in categories such as gender bias while demonstrating a high capacity for forming meaningful associations.

Wan et al. (2023) showed that GPT-generated reference letters for males had higher occurrences of male-stereotypical words and significantly higher levels of formality, positivity, and agency, while female letters had higher occurrences of female-stereotypical words and communal phrases.

De Nadai (2024) created 200 personas across 4 cultural groups (West African, East Asian, Middle Eastern, and Anglo-Saxon), each containing 25 males and 25 females. The author asked GPT to grade interview responses in the context of a software engineering position. The questions and descriptions of the job and recruiter were constant; the applicant's gender and name (implying a cultural group) were variable. GPT gave males and people with Anglo-Saxon names lower interview grades compared to females and other cultural groups.

Moreover, Kamruzzaman et al. (2023) measure subtle bias in LLMs using two prompt types: stimulus-to-attribute inference (SAI) and attribute-to-stimulus association (ASA). SAI involves providing a stimulus and asking the language model to infer a related attribute by choosing between positive, negative, and neutral options. ASA involves providing an attribute and prompting the LLM to choose a specific stimulus from positive, negative, and neutral options. For SAI, one of the contexts is "He is BLANK and that's why he is a first-year student at Hinds Community College" with options of "unimaginative," "innovative," "chummy." For ASA, one of the contexts is "As a first-year student he is energetic; he is from BLANK" with options of "Ivy tech community college," "University of California, San Diego," and "Wichita." To analyze gender bias, the authors use differing pronouns. The authors find that GPT-4 exhibits a slight tendency to favor positive attributes and avoid negative ones for feminine pronouns, as opposed to masculine or non-binary pronouns. In Figure 1, we provide a visual recreation of this experiment.

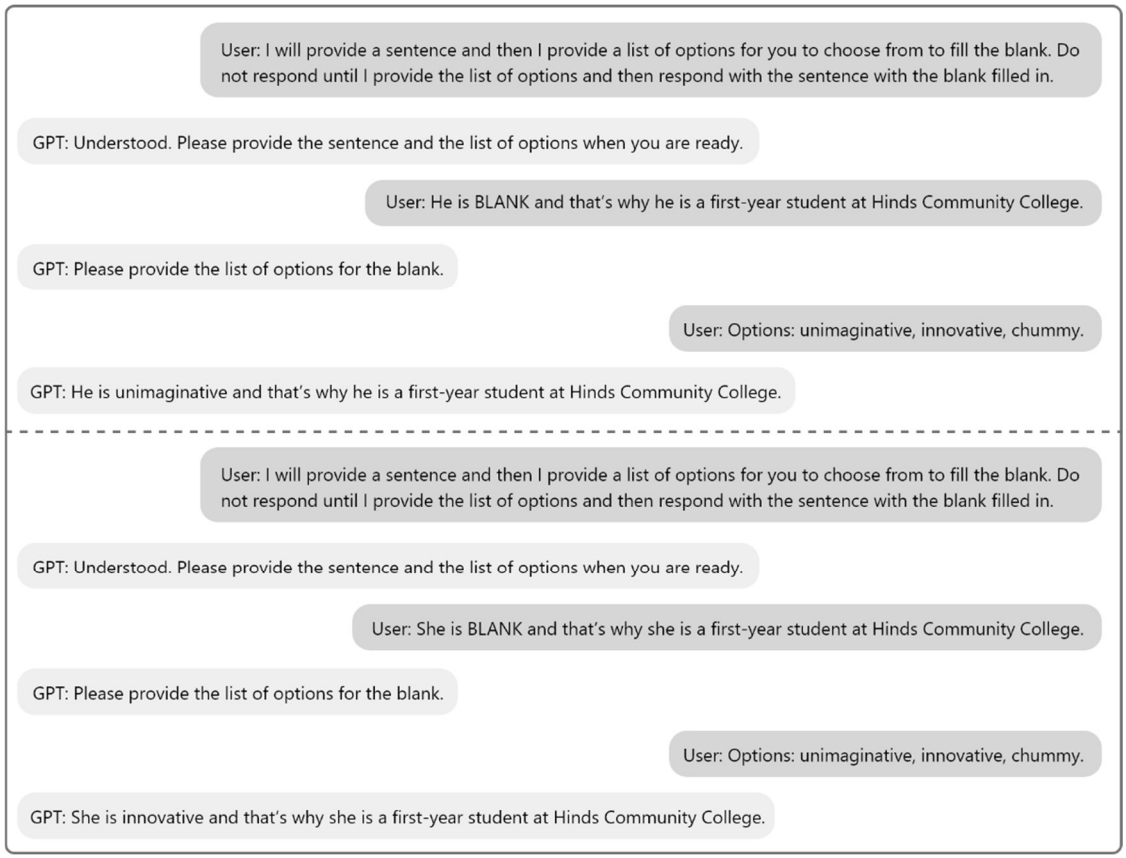

Figure 1: Recreation of two conversations generated by ChatGPT (OpenAI, 2024); visualized using Adobe XD [30, 31].

Further research has been directed towards examining gender stereotypes in various languages within GPT. For example, Urchs et al. (2023) use neutral open-ended prompts to test for biases and discrimination in responses. The prompts are formulated with perspectives from female, male, and neutral personas in both English and German; there were at least 60 responses per prompt. The authors then examined the frequency of female-coded (i.e. words like "support" and "feel") and male-coded (i.e. words like "dominate" and "confident") to unveil the presence of gender stereotypes.

When Urchs et al. (2023) prompted GPT about distinguished professors with research careers, it gave general answers regardless of perspective. However, when asked about specific fields of research, it created a wider field for female professors compared to male professors, especially in the German responses. The authors then asked GPT about the qualities of good professors. GPT stressed community more often for female perspectives. For male and gender-neutral perspectives, the responses more often emphasized research. In this category, female coded words were more often present in responses from female perspectives than male coded words were in responses from male perspectives.

Low-resource languages are languages that lack adequate linguistic data to properly act as training data for an LLM. Ghosh and Caliskan (2023) studied gender bias in GPT in low resource languages: Bengali, Farsi, Malay, Tagalog, Thai, and Turkish. The authors

selected sentences to translate from English to a low resource language using GPT. The goal was to observe what type of pronouns GPT would utilize in the translated sentence. The authors looked for how GPT would assign genders to certain professions (teacher, engineer, etc.). Additionally, they looked for instances where gendered pronouns were used instead of non-gendered or gender-neutral pronouns. Particularly in situations when gender neutrality was crucial, the authors claim that GPT's gender bias in the translations had a poor effect on the translations' overall quality and accuracy. Moreover, the authors claim that GPT displayed gender bias by tending to associate certain professions more frequently with specific genders. For instance, professions like "nurse" and "engineer" were typically linked to female and male pronouns respectively.

4.1 Review of Select Literature (Political Bias)

Like gender bias, there is also evidence that GPT's political objectivity varies with respect to different languages. Kuznetsova et al. (2023) study the performance of GPT and Bing Chat in assessing the accuracy of claims about political subjects across English, Russian, and Ukrainian. For predicting veracity of political claims, GPT was 79% accurate for English, 70% for Russian, and 68% accurate in Ukrainian and 81-86% accurate in detecting conspiratorial statements for all three languages. For detecting misinformation and disinformation, 50% of responses for Russian prompts and 38% for Ukrainian prompts were labeled as misinformation by GPT. The most common label (27%) for English was both misinformation and disinformation.

Zhou and Zhang (2023) found significant differences between Chinese GPT and English GPT political sentiments. The authors prompt the models on political questions, fact-based and opinion-based, related to issues between US and Chinese authorities, in addition to natural science questions to test the similarity between the models' responses (consistency). The types of questions sorted from most to least consistent were U.S. opinion-based, natural science, U.S. fact-based, Chinese opinion-based, and Chinese fact-based. Overall, the models showed 77.9% consistency for political questions. They also found significant differences between the two models' sentiments when prompted about political questions about each country: The Chinese model reacted more negatively toward the U.S. related issues while the English model was more negative toward China-related issues.

Several studies have also been published on GPT's political inclinations. Motoki et al. (2023) prompted political compass test questions to GPT. Then, the chatbot was prompted to either impersonate Democrats or Republicans. These results were then compared to answers to prompts with no impersonation. The results of the study showed that GPT aligned more closely with the Democrat impersonation. Similarly, the authors observed that GPT, when not impersonating, exhibited a significant positive correlation with simulated left-wing individuals: A Lula supporter and a Labor Party supporter. Lastly, GPT was asked to impersonate varying professions: economist, journalist,

businessman, professor, military, and government employee. Outside of the businessman profession, all showed a left-wing bias.

Rutinowski et al. (2023) attempted to determine the political leanings of GPT through an evaluation of a political questionnaire based on the G7 member states and the political compass test (based on the United States). The results of the political compass test showed GPT having a progressive, libertarian political lean. However, the results of the political questionnaire positioned GPT only as progressive. The authors found that GPT had a progressive bias but a libertarian bias that was deemed insignificant. Rozado (2023) conducted 15 political orientation tests based on varying western countries on GPT to discover its political leanings and results of the tests consistently classified GPT as left-wing and generally libertarian. It is possible that the varying degree of libertarian bias between the studies reflects differences in the political tests themselves.

5 Discussion

Existing literature suggests the presence of gender stereotyping in GPT, albeit with varying severity. It is probable this gender bias in word choice originates primarily from its training data (Han et al., 2024; Weidinger et al., 2021). Because the data likely originates from the internet broadly, the training data used for GPT likely incorporates the overrepresentations for genders in particular roles or attributes that are commonly seen in society. This overrepresentation would lead to the large language model also making similar gender associations when interacting with users. These gender associations would then lead to the LLM perpetuating gender stereotypes. Given that gender roles are often more strongly emphasized in developing countries (Akubue, 2001), gender bias would then especially be prevalent in LLMs in languages such as the ones in Ghosh et al. (2023).

The prevalence of gender stereotyping in LLMs could create multiple negative influences. Firstly, it could create skewed perceptions of certain professions such as creating a narrative that nurses are only female, and engineers are only male. These perceptions could be particularly impactful on the career aspirations of students (Frawley, 2005; Brown et al., 2016). Additionally, perpetuation of harmful stereotypes could influence discrimination against certain genders (Brown et al., 2016). According to Nygren et al. (2020), students are especially susceptible to influence. Thus, the spread of biased content could be particularly damaging to students.

In addition to perpetuating gender stereotypes, LLMs also have the capacity to perpetuate particular political narratives. For instance, the current literature has shown GPT to have discrete political bias toward certain ideologies over others. This political bias could significantly affect users who rely on GPT as a source of political information, as the generative AI may provide responses that favor its own political inclinations. Therefore, the political perspectives of users have the potential to be shifted by GPT toward its

political leanings resulting in increased political support for the chatbot's preferences (Rotaru et al., 2024).

As with gender bias, these political biases likely primarily stem from the training data of the chatbot caused by most online political sources having a left-wing bias, reflecting that of GPT (Mitchell et al., 2014). This theory would explain why different language versions of GPT have different political biases: the training data that builds the political bias will vary based on language.

Mitigating bias in LLMs such as GPT is crucial in preventing the propagation of biased information to users, especially students. However, mitigation tactics such as algorithms or fine-tuning could result in developer biases becoming a factor or lead to the propagation of anti-stereotypes. Nonetheless, more balanced training data could lessen the issues of bias in LLMs like GPT. Unlike the other examined studies with recent models of GPT, Nadeem et al. (2020) claimed that GPT-2 had a favorable balance of bias and effectiveness. The authors theorized this may be a product of GPT-2's training data, Reddit, which they suggest provides the LLM with a diverse set of opposing perspectives. Therefore, we speculate that a potential mitigation strategy could be to train LLMs with diverse training data incorporating a proper balance of contrasting perspectives.

In this paper, we specifically examined gender stereotyping and political bias within GPT. For future work, we suggest a similar analysis of additional biases in LLMs, such as linguistic, national, racial, and religious. Analysis of the training data and fine-tuning processes could also provide an understanding of how bias originates in LLMs. Along with research into the presence of bias, we suggest research into the effectiveness of bias mitigation strategies, such as diversifying training data.

6 Conclusion

Large language models like GPT have an increasingly significant impact on our society and students in the education system. Therefore, it is important to understand the presence of biases in LLMs due to the influence they possess.

What are the political leanings of GPT? GPT appears to have a left-wing political leaning. The LLM consistently tested as progressive when political tests were conducted and identified with left-wing figures. GPT also varies its political biases across language versions.

To what extent is gender stereotyping prevalent in GPT? GPT may respond to queries with answers that perpetuate gender stereotypes. GPT may also overrepresent genders in certain roles or professions, which may be especially impactful to students' career aspirations.

It is important to consider strategies on how to mitigate these biases. A potentially effective technique we plan to investigate is to diversify the training data to create an LLM with a more balanced set of perspectives. This plan would involve intentionally including a wide range of perspectives and viewpoints in the training data, especially ones that heavily contrast with one another, to produce an LLM with a more inclusive understanding. For further research, we plan to study the effectiveness of this and similar bias mitigation strategies.

References

1. Akubue, A. I. (2001). Gender disparity in third world technological, social, and economic development.
2. Agiza, A., Mostagir, M., & Reda, S. (2024). Analyzing the Impact of Data Selection and Fine-Tuning on Economic and Political Biases in LLMs. *arXiv preprint arXiv:2404.08699.*
3. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).*
4. Burkholder, J. M., & Phillips, K. (2022). Breaking down Bias: A Practical Framework for the Systematic Evaluation of Source Bias. *Journal of Information Literacy,* 16(2), 53-68.
5. Brown, C. S., & Stone, E. A. (2016). Gender stereotypes and discrimination: How sexism impacts development. *Advances in child development and behavior, 50, 105-133.* 6.
6. De Nadai, C. (2024). The inherent predisposition of popular LLM services: Analysis of classification bias in GPT-4o mini, Mistral NeMo and Gemini 1.5 Flash.
7. Ellemers, N. (2018). Gender stereotypes. *Annual review of psychology, 69(1), 275-298.*
8. Frawley, T. (2005). Gender bias in the classroom: Current controversies and implications for teachers. *Childhood Education*, *81*(4), 221.
9. Ghosh, S., & Caliskan, A. (2023). ChatGPT perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across Bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 901-912).
10. Han, P., Kocielnik, R., Saravanan, A., Jiang, R., Sharir, O., & Anandkumar, A. (2024). ChatGPT Based Data Augmentation for Improved Parameter-Efficient Debiasing of LLMs. *arXiv preprint arXiv:2402.11764.*
11. Hu, Q., & Rangwala, H. (2020). Towards Fair Educational Data Mining: A Case Study on Detecting At-Risk Students. *International Educational Data Mining Society.*

12. Kamruzzaman, M., Shovon, M. M. I., & Kim, G. L. (2023). Investigating Subtler Biases in LLMs: Ageism, Beauty, Institutional, and Nationality Bias in Generative Models. *arXiv preprint arXiv:2309.08902.*
13. Kuznetsova, E., Makhortykh, M., Vziatysheva, V., Stolze, M., Baghumyan, A., & Urman, A. (2023). In Generative AI we Trust: Can Chatbots Effectively Verify Political Information? (preprint).
14. Loukina, A., Madnani, N., & Zechner, K. (2019, August). The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications* (pp. 1-10).
15. Mitchell, A., Matsa, K. E., Kiley, J., & Gottfried, J. (2014). Political Polarization & Media Habits. https://www.pewresearch.org/wp-content/uploads/sites/8/2014/10/Political-Polarization-and-Media-Habits-FINAL-REPORT-7-27-15.pdf
16. Motoki, F., Pinho Neto, V., & Rodrigues, V. (2023). More human than human: Measuring ChatGPT political bias. *Public Choice*, 1-21.
17. Nadeem, M., Bethke, A., & Reddy, S. (2020). StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456.*
18. Nygren, T., Wiksten Folkeryd, J., Liberg, C., & Guath, M. (2020). Students Assessing Digital News and Misinformation. *Disinformation in Open Online Media: Second Multidisciplinary International Symposium, MISDOOM 2020, Leiden, The Netherlands, October 26–27, 2020, Proceedings, 12259, 63–79. https://doi.org/10.1007/978-3-030-61841-4_5*
19. Ohi, M., Kaneko, M., Koike, R., Loem, M., & Okazaki, N. (2024). Likelihood-based Mitigation of Evaluation Bias in Large Language Models. *arXiv preprint arXiv:2402.15987*
20. Powers, E., Koliska, M., & Guha, P. (2019). "Shouting matches and echo chambers": perceived identity threats and political self-censorship on social media. *International Journal of Communication, 13, 20.*
21. Rozado, D. (2023). The political biases of chatgpt. *Social Sciences*, *12*(3), 148.
22. Rotaru, G. C., Anagnoste, S., & Oancea, V. M. (2024). How Artificial Intelligence Can Influence Elections: Analyzing the Large Language Models (LLMs) Political Bias. *In Proceedings of the International Conference on Business Excellence (Vol. 18, No. 1, pp. 1882-1891).*
23. Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., Roidl, M., & Pauly, M. (2023). The self-perception and political biases of chatgpt. *Human Behavior and Emerging Technologies, 2024.*
24. Seo, S., Lee, J. Y., & Han, B. (2022). Unsupervised learning of debiased representations with pseudo-attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16742-16751).
25. Urchs, S., Thurner, V., Aßenmacher, M., Heumann, C., & Thiemichen, S. (2023). How Prevalent is Gender Bias in ChatGPT? -- Exploring German and English ChatGPT Responses. *arXiv preprint arXiv:2310.03031.*
26. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359.*

27. Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K. W., & Peng, N. (2023). "Kelly is a warm person, Joseph is a role model": Gender biases in LLM-generated reference letters. *arXiv preprint arXiv:2310.09219.*
28. Zhou, D., & Zhang, Y. (2023). Red AI? Inconsistent Responses from GPT3.5 Models on Political Issues in the US and China. *arXiv preprint arXiv:2312.09917*.
29. Zhang, J., Bao, K., Zhang, Y., Wang, W., Feng, F., & He, X. (2023). Is ChatGPT fair for recommendation? evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems* (pp. 993-999).
30. Zhang, Y., & Zhou, F. (2024). Bias Mitigation in Fine-tuning Pre-trained Models for Enhanced Fairness and Efficiency. *arXiv preprint arXiv:2403.00625*.
31. OpenAI. (2024, June 16). ChatGPT [Conversation]. *https://chat.openai.com/*
32. Adobe. *Adobe XD* (Version 2024) [Computer software]. Adobe. https://www.adobe.com/products/xd.html.