Testing LLM’s Ability in Mitigating Biases

Sophie Hsiao, Fiona Fang, Hugo Hsieh, Jianfeng Chen

Abstract

This research aims to investigate whether Large Language Models (LLMs), can mitigate biases in textual content. Using a dataset comprising biographical information, the research initially quantifies gender biases in the profession using word embedding techniques. Then, LLMs like Google Gemini and ChatGPT-3.5 are tasked with rewriting the text, followed by recalculating bias scores for comparisons. The findings show that with the intervention of prompt in LLM, a certain degree of bias can be mitigated, though with very limited effect. Through the experimental approach, this study aims to provide insights into the potential of LLMs as tools for promoting fairness and equity in textual communication, thereby contributing to the broader discourse on mitigating societal biases.

1 | Introduction

The advancement of Large Language Models (LLMs) has revolutionized natural language processing (NLP) capabilities, empowering various applications across domains such as text generation, sentiment analysis, and machine translation. However, amidst their transformative potential, concerns have arisen regarding the potential biases within LLM-generated text. Biases, whether implicit or explicit, can manifest in various forms, including gender stereotypes, racial prejudices, and occupational biases, thereby perpetuating societal inequities and reinforcing discriminatory norms.

Recognizing the imperative to mitigate such biases, this study delves into the potential of LLMs in detecting and rectifying biases within textual content. There are two types of bias when employing LLMs. The first, intrinsic bias, refers to biases inherently embedded within the representations generated by the pre-trained model. These biases originate from the data sources utilized during model training and are independent of any specific downstream tasks. Conversely, extrinsic bias, the second type, becomes evident in the model output during downstream tasks, such as classification or generation tasks. This form of bias is tied to the nature and requirements of the particular task being performed.

Our study recognizes the existence of the above-mentioned biases, aiming to see how they are shown in LLMs and examining the effectiveness of LLMs, particularly Google Gemini and ChatGPT-3.5, in addressing and correcting them. We are interested in the following research question:

R1*: How effectively can LLMs reduce biases present in the text?*

What's more, different LLMs operate differently due to variations in their architecture, training data, and fine-tuning processes. These differences can impact their ability to address biases. Therefore, understanding the comparative performance of various LLMs is essential for developing strategies for bias mitigation. In this context, we propose the second research question:

R2: *Which LLM is more effective at mitigating biases present in the text?*

By quantifying and examining the capabilities of LLMs in mitigating biases, this research aspires to contribute to the ongoing discourse on bias mitigation in AI technologies and pave the way for more equitable and socially responsible applications of natural language processing.

2 | Literature Review

The existing literature demonstrates that LLMs themselves exhibit societal biases. In a recent investigation by Kotek et al. (2023), the behavior of LLMs regarding gender stereotypes was explored. Testing four recent LLMs, the study reveals a tendency for these models to manifest biased assumptions about men's and women's occupations, often favoring stereotypical choices. LLMs are found to be 3-6 times more likely to select occupations stereotypically aligned with a person's gender. Additionally, Fang et al. (2023) explored biases beyond word analysis, examining disparities in AI-generated content by LLMs compared to reputable sources like The New York Times and Reuters. Their findings suggest that LLM-generated content may convey sentiments and toxicities toward various gender or race-related population groups, demonstrating discrimination against underrepresented communities.

Moreover, variations in bias among different LLMs have been observed, as evidenced by a recent study comparing the performance of ChatGPT-4, Gemini, and Llama 2 (Chan & Wong, 2024). ChatGPT-4 exhibited higher biases in gender, race, and ethnicity representation, while Llama 2 showed the least bias overall. Also, Llama 2 achieves the highest fairness scores. This suggests that Llama 2's methods for correcting bias are more effective compared to the other models, with Google Gemini performing the worst.

Various techniques have been proposed to detect and mitigate biases in LLMs. In a study by Dwivedi et al. (2023), innovative approaches such as prompt engineering and in-context learning are introduced to rectify biases in LLMs. Through these methods, LLMs are guided to produce more equitable content, resulting in a reduction in gender bias, particularly in areas traditionally prone to biases like 'Literature'. Similarly, research by Cai et al. (2024) demonstrates that the LSDM (Least Square Debias Method), a knowledge-editing-based method for mitigating gender bias in occupational pronouns, can help mitigate bias while preserving the LLMs' capabilities.

In summary, while recent research demonstrates promising advancements in mitigating biases within LLMs, challenges remain in addressing inherent societal biases. Building upon these foundational studies, our research aims to delve deeper into assessing the effectiveness of different LLMs on "cleansing" gender biases within professions in the text.

3 | Method

Figure 1 provides a general overview of the data analysis process. The research unfolds across four primary stages: data collection and preprocessing, bias score computation, LLM intervention to correct the text, and bias score recalculation for comparative analysis.

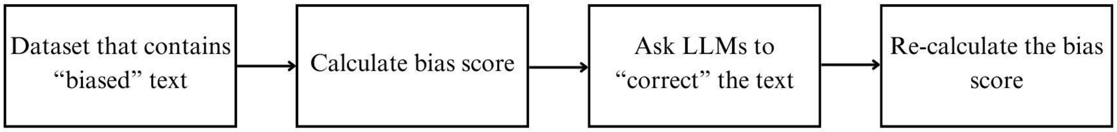

Dataset that contains "biased" text
Calculate bias score
Ask LLMs to "correct" the text
Re-calculate the bias score

Fig.1 Research Method

3.1 Data

This study utilizes the Bias in Bios dataset[1], which was designed to evaluate bias in NLP models. Developed by De-Artega et al. in 2019, this dataset consists of textual biographies alongside associated professional occupations and gender attributes. The dataset is divided between train (257,000 samples), test (99,000 samples), and dev (40,000 samples) sets. Our study mainly utilizes the dev sets.

3.2 Word Embeddings for Calculating Bias

Traditional bias measurement methods often contain dictionary-based or qualitative approaches, which can be time-consuming and challenging to apply across different languages, temporal spans, and stereotype categories. Word embeddings, a technique that represents words as vectors through deep neural networks, offer a solution by revealing semantic relationships based on their positions in space. As similar words are closer in their space, word embedding can theoretically identify biases. Caliskan, Bryson, and Narayanan (2017) and Garg, Schiebinger et al. (2018) have already proposed how to use word embedding as a tool to measure bias and validated such methods by correlating bias indices calculated through word embedding with gender and ethnicity compositions across various U.S. Census occupations.

To capture the semantic relationship of words in mathematical representations, we adopt the Continuous Bag of Words (CBOW) model. The CBOW model is a predictive algorithm utilized within Natural Language Processing to hypothesize the probability of a word given a context—a sequence of surrounding words. Its architecture is designed to predict the missing center word

[1] The dataset can be accessed on https://huggingface.co/datasets/LabHC/bias_in_bios.

based on the context provided by neighboring words. For instance, given the sentence “DSPP is definitely a good course to take…”, the model would aim to predict ‘definitely’ when given the context words “DSPP, is, a, good.” In the effort to make predictions accurate, the model would learn the weights that reflect semantic relationships between words well.

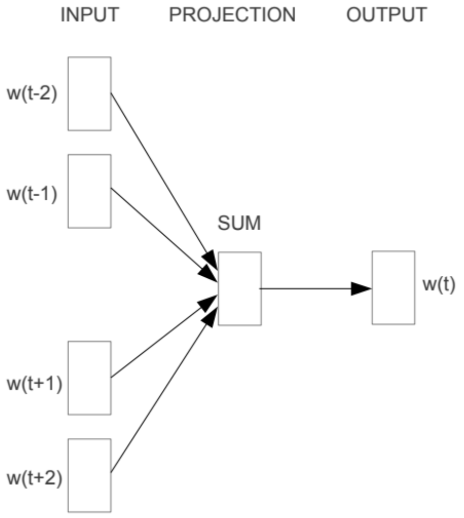


Fig. 2 Continuous Bag-of-Words (CBOW) Model Architecture

Therefore, the weights obtained from training with the CBOW model are used as the word embeddings, which transform the lexical items into points within a semantic space. For example, the words ‘nurse’ and ‘boy’ could be represented by vectors [2, 3, 4, 6, 5, 3, ...] and [3, 2, 4, 8, 4, 5, ...], respectively. This vector representation enables us to calculate similarity in the semantic space, which serves as a proxy for semantic similarity and, by extension, biases.

In practice, the vectors for gender-related words are averaged to create a centroid vector that typifies the gendered concept within the semantic space. The distance—or in our case, the cosine similarity—between this centroid and the profession vector is then calculated. We retrieve the vectors for gender-related words and the target profession, compute the average embeddings for each gender, and then calculate the cosine similarities. These similarities are reported in the output, revealing the relative gender bias for the profession in question.

Specifically, to quantify gender bias, we map gender-related concepts to their corresponding vectors. We aggregate vectors representing female-related words such as 'woman,' 'girl,' and 'she,' and similarly for male-related words like 'man,' 'boy,' and 'he.' By computing the cosine similarity—a measure that quantifies the similarity between two vectors—we can discern the proximity between gendered word sets and a given profession. For instance, if the cosine similarity between 'nurse' and female-associated words is greater than that between 'nurse' and male-associated words, this suggests a female bias for the profession of nursing.

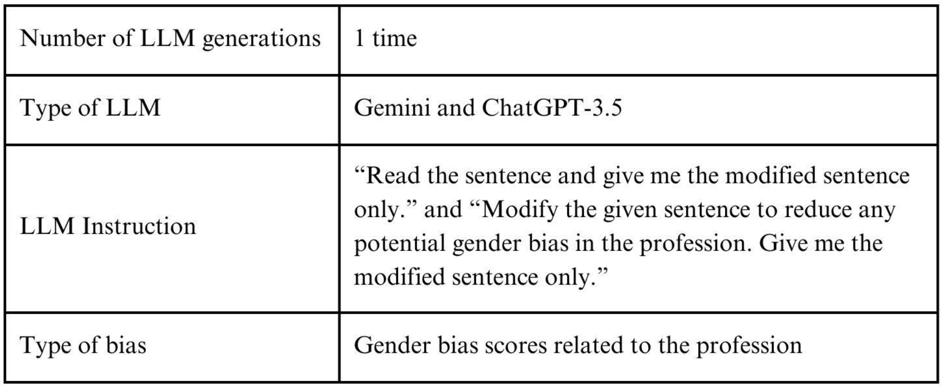
By leveraging the CBOW model in this way, our study adopts an empirical, data-driven approach to quantifying biases that may exist in language models, thus allowing for a systematic assessment of gender biases in textual content.

3.3 LLMs Experimental Design

To investigate the effectiveness of LLMs in correcting biases, this study develops an experimental design. Before experiments, gender bias scores in the profession of the original texts will be calculated and recorded using the CBOW model.

The experimental scenarios will be divided based on “Number of LLM generations,” “Type of LLM,” “LLM Instruction,” and “Type of bias.” Specifically, “Number of LLM generations” refers to the number of times the original text is rewritten using an LLM, only 1 time in this study; “Type of LLM” refers to the LLM product and version used, including Gemini and

ChatGPT-3.5; "LLM Instruction" refers to the commands given to the LLM for generation, including "Read the sentence and give me the modified sentence only." and "Modify the given sentence to reduce any potential gender bias in the profession. Give me the modified sentence only."; and "Type of bias" includes gender bias scores related to the profession.



Number of LLM generations	1 time
Type of LLM	Gemini and ChatGPT-3.5
LLM Instruction	"Read the sentence and give me the modified sentence only." and "Modify the given sentence to reduce any potential gender bias in the profession. Give me the modified sentence only."
Type of bias	Gender bias scores related to the profession

Fig 3. Experimental Design

This study utilizes Google Gemini and ChatGPT-3.5 as testing models. Google Gemini, equipped with a free-access API, presents a cost-effective option for research endeavors. Additionally, Gemini offers many different models, such as Gemini Ultra, Nano, and Pro, ensuring adaptability to diverse research needs and situations. On the other hand, ChatGPT-3.5 stands out for its affordability. With its cost-effective API, we can leverage its capabilities without incurring significant costs.

This experimental design allows us to extend beyond examining differences in biases under singular conditions. We explore the effects across different experimental setups, including different LLMs and under passive and proactive instructions, to assess their efficacy in "cleansing" biases.

4 | Results

Figure 3 presents the gender tendencies across various professional domains, comparing the original text with two versions of the text rewritten by LLMs, each with and without prompts. In Figure 4, a heatmap illustrates the gender bias tendencies score.

In the original text, distinct gender biases are evident: nurses, journalists, and dietitians exhibit positive values, indicating a bias towards females. Conversely, engineers and professors demonstrate a significant negative value, suggesting a bias towards males. Other professions, such as doctors, teachers, and dentists, exhibit values closer to neutral, indicating a less pronounced gender bias in the original text.

In the nursing sector, the original text exhibits a pronounced female bias, which was notably diminished when utilizing prompts in the context of the Gemini model. However, this effect of prompt inclusion did not manifest as significantly within the ChatGPT-3.5 model. As for the medical profession, specifically doctors, the Gemini model initially revealed a slight female bias, which subtly transitioned to a male bias upon the integration of prompts. The ChatGPT-3.5 model demonstrates a relatively stable male bias that appeared relatively impervious to the presence or absence of prompts.

Delving into education-related professions such as teaching and academia, the original text was skewed towards female bias, which was lessened with the application of prompts in the Gemini model. In a rather contrasting display, the ChatGPT 3.5 model tended towards a male bias, which was more pronounced when subjected to prompts, indicating a differential response to prompt-based adjustments.

Technical and communicative fields presented their unique bias patterns. The engineering profession, as depicted by the original text, was strongly biased towards males, and this bias remained largely unchanged irrespective of prompts in both the Gemini and ChatGPT-3.5 models. On the other hand, journalism saw a shift towards female bias in the Gemini model when prompts were introduced, a shift not paralleled in the ChatGPT-3.5 model's behavior.

Health and wellness professions also conveyed biases, albeit more subtly. The dietitian field showed a consistent female bias across all models and conditions. In contrast, the dentistry field, which the original text associated more with males, experienced a slight reduction in bias with the introduction of prompts in the Gemini model. The ChatGPT-3.5 model, when prompted, interestingly indicated a reversal of this bias.

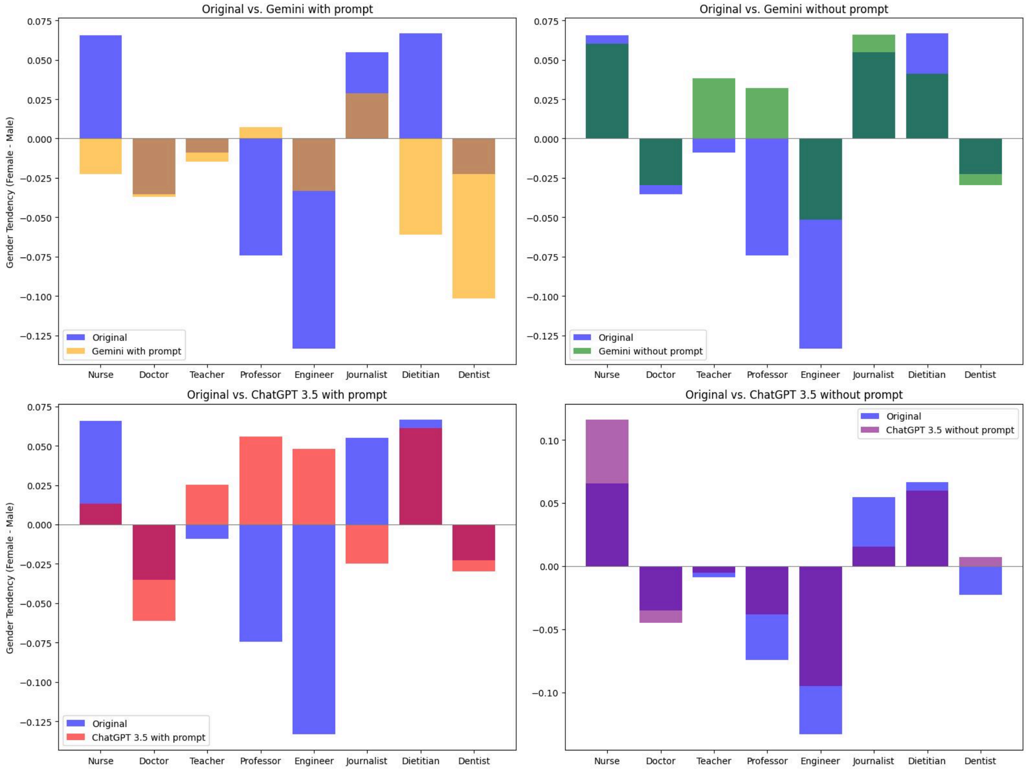


Fig 4. Gender Tendency between Original Text and Rewritten Text by LLMs

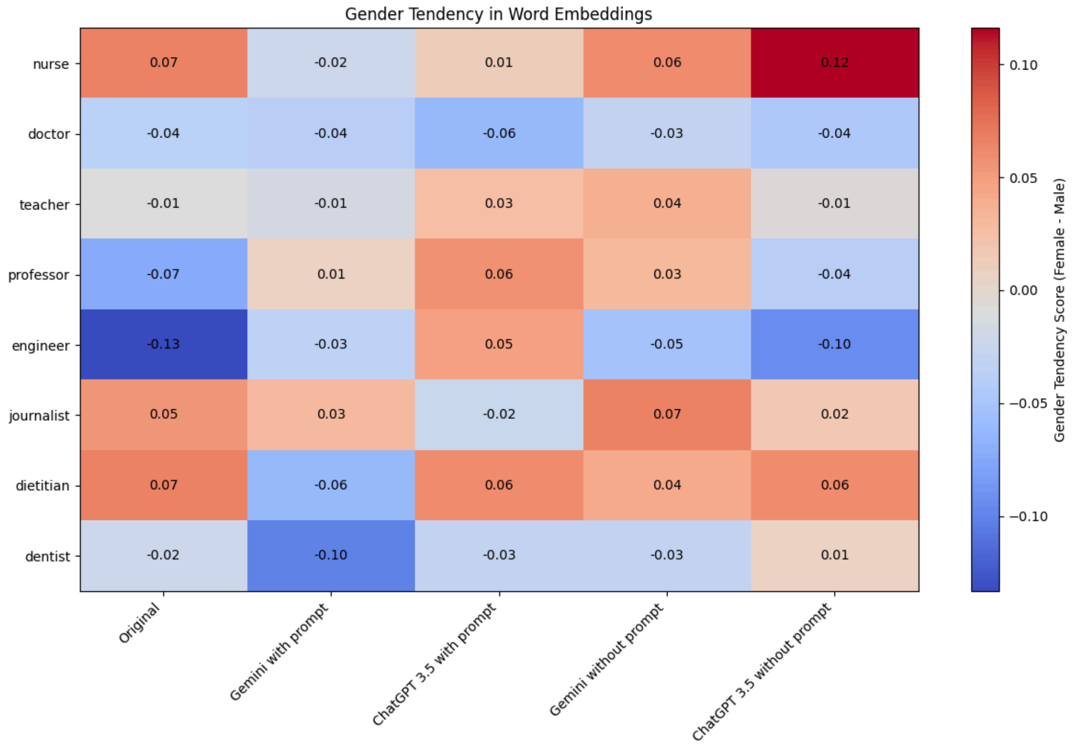


Fig 5. Gender Tendency in Heatmap

Overall, the findings indicate that while prompts can modify the word embeddings towards gender bias in the text, the direction and magnitude of the effect vary not only with the profession but also with the specific LLM. This underscores the complexity of mitigating biases in LLMs and suggests that interventions must be carefully designed and tailored to the intricacies of each model and context.

5 | Discussion

The unprecedented fast development of LLM poses great challenges for society and regulators. In 2023, the Biden Administration secured voluntary commitments from LLM developers like Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI to help move the development of AI technology toward safe, secure, and transparent (The White House, 2023). However, the development of policy regarding fairness and bias mitigation in each developer remains incoherent and slow. Taking ChatGPT and Gemini as case studies, in this section we will further look into their respective policy toward fairness.

As the pioneer in the LLM industry, OpenAI was accused of cases with biased content in its LLM model - ChatGPT. In addressing the bias and flaws of its model, OpenAI provides a simple one-page policy and explanation on how it is developing relevant methods to mitigate (Open AI, 2023). Mostly focusing on using reviewers to monitor and techniques such as reinforcement learning from human feedback (RLHF), OpenAI recognized the flaws and biases in its system but has rather limited research and dedicated personnel in furthering the process. In addition, OpenAI did not have a developed policy on its AI's ethical and fairness principle, but only called for the public's assistance in furthering the research (OpenAI, 2023). Looking specifically at gender, OpenAI does not provide any information on the mitigation of gender bias policy.

Being the major player in the industry for ages, Google has a rather more thorough policy regarding the development of AI. Google's AI Principles serve as the backbone of its responsible AI development, with emphasis on the AI application being socially beneficial, mitigating unfair biases, being safe, accountable, retaining privacy, and reaching scientific excellence (Google, n.d.). In dealing with the fairness of AI, Google provides frameworks to guide developers and implementers to ensure fairness. Numerous strategies have been proposed to promote fairness and inclusivity in AI development and deployment. One key approach involves building a diverse and inclusive workforce, ensuring that a wide range of perspectives contribute to AI projects. Engaging with communities early in the research process is also crucial for understanding societal contexts. Additionally, thorough examination of training data for bias, training models to remove biases, evaluating model performance, and ongoing testing for fairness are essential steps in creating AI systems that are equitable and ethical (Google, n.d.). Looking specifically at gender bias mitigation, Google recognized the existence of gender biases, especially in professional settings, and provided guidance and research to help further the research on this domain (Google, n.d.).

Examining the policies of prominent LLM developers, Google's framework stands out for offering clearer guidance to LLM users and researchers. It integrates various techniques along with social and scientific research to address bias more effectively. OpenAI, on the other hand, lacks comprehensive policy and dedicated team albeit being the pioneer in developing LLM models. Our research findings also indicate that Google Gemini demonstrates superior mitigation capabilities compared to ChatGPT regarding mitigating gender bias in professions. Upon examining their respective policies, it's evident that Google has invested more effort in crafting comprehensive frameworks and policies for mitigation, which are reflected in their models. While it's challenging to definitively conclude that a more developed policy framework directly leads to increased fairness in LLM models, it does signify the level of effort developers invest in addressing bias.

Based on our research findings, we propose that both government entities and the public act as overseers to incentivize prominent LLM developers to establish a "coherent, tangible, and comprehensive" set of principles and frameworks as initial steps. By fostering collaboration between the public and private sectors and implementing effective monitoring mechanisms, such a framework could serve as a consensus for future regulatory efforts.

6 | Conclusion

In summary, this research explored the potential of LLMs, specifically Google Gemini and ChatGPT-3.5, in mitigating gender bias within textual content related to professions. While both models demonstrated some ability to reduce bias, the effectiveness varied depending on the specific LLM as well as the presence of prompts. Our findings include:

- LLMs can mitigate bias while limitations exist: Both Gemini and ChatGPT-3.5 showed the potential to reduce gender bias, particularly when prompts were used. However, the impact was not uniform across all professions, and some biases remained persistent.
- Performance varies between LLMs: Gemini exhibited a greater ability to mitigate bias compared to ChatGPT-3.5, aligning with the observation that Google has invested more in developing comprehensive frameworks and policies for bias mitigation.
- Prompts play a crucial role: The inclusion of prompts significantly influenced the results, highlighting the importance of carefully designed instructions in guiding LLMs towards fairer outputs.

Our research suggests there's more to explore in the future. First, further investigation of the effectiveness of different models, such as Llama 2, in mitigating various biases (e.g., racial, age-related) is crucial. Second, refining prompt engineering techniques is essential for developing more effective prompts to maximize the positive impact of LLMs. Third, addressing intrinsic bias by investigating methods to tackle biases in the training data and model architecture is crucial for long-term solutions. Lastly, developing comprehensive policy frameworks through collaboration between LLM developers, researchers, and policymakers is necessary to establish clear guidelines and standards for responsible AI development.

This research contributes to the ongoing discourse on mitigating bias in AI and highlights the potential of LLMs as tools for promoting fairness and equity in textual communication. As LLM technology continues to evolve, ongoing research and responsible development practices are necessary to ensure these powerful tools are used ethically and contribute to a more just and equitable society.

References

Cai, Y., Cao, D., Guo, R., Wen, Y., Liu, G., & Chen, E. (2024). Locating and Mitigating Gender Bias in Large Language Models. https://doi.org/10.48550/arXiv.2403.14409

Chan, M., & Wong, S. (2024). A Comparative Analysis to Evaluate Bias and Fairness Across Large Language Models with Benchmarks. https://doi.org/10.31219/osf.io/mc762

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

De-Arteaga et al. (2019). Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 120–128). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287572

Dwivedi, S., Ghosh, S., & Dwivedi, S. (2023). Breaking the Bias: Gender Fairness in LLMs Using Prompt Engineering and In-Context Learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities, 15*(4).

Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., & Zhao, X. (2023). Bias of AI-generated content: an examination of news produced by large language models. *Scientific Reports volume 14***(**5224). https://doi.org/10.1038/s41598-024-55686-2

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018, April 3). Word embeddings quantify 100 years of gender and ethnic stereotypes. In *Proceedings of the National Academy of Sciences*, *115*(16). https://doi.org/10.1073/pnas.1720347115

Google. (n.d.). *Our Principles.* Google. https://ai.google/responsibility/principles/

Google. (n.d.). *Responsible AI practices.* Google. https://ai.google/responsibility/responsible-ai-practices/

Google. (n.d.). *Safety & Fairness Considerations for Generative Models.* Google. https://developers.google.com/machine-learning/resources/safety-gen-ai

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference* (12–24). Association for Computing Machinery. https://doi.org/10.1145/3582269.3615599

Open AI. (2023, Feb, 16). *How should AI systems behave, and who should decide?* Open AI. https://openai.com/blog/how-should-ai-systems-behave

The United States Government. (2023, July 21). *Fact sheet: Biden-Harris Administration secures voluntary commitments from leading artificial intelligence companies to manage the risks posed by AI.* The White House. https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

Collaboration Statement

Name	Contributions
Hugo Hsieh	Responsible for the CBOW model construction, training, word embedding extraction, bias measurement, and visualization.
Fiona Fang	Lead the report writing - Plan the overall report structure and format - Write the Abstract, Introduction, Literature Review, and Method parts - Revise the Results, Conclusion, and Reference parts
Sophie Hsiao	Lead the report writing - Discussion on Bias - Investigate LLM developers' policy and suggestion
Jianfeng Chen	Responsible for prompt engineering, data generation by Gemini and Gpt 3.5, data visualizations and result analysis.