Research Proposal: Testing LLM's Ability in Mitigating Biases

This project aims to investigate the potential of widely used Large Language Models (LLMs), such as ChatGPT, in mitigating stereotypes and biases within text. Specifically, we seek to assess the capability of LLMs to identify and rectify biased language when presented with such text. By systematically examining the performance of LLMs in a biased corpus under various experimental settings, this study will contribute by addressing the gap in understanding the efficacy of LLMs in correcting societal biases and the potential of LLMs as tools for promoting fairness and equity.

Literature shows that some measures can be taken to improve LLMs' capabilities in mitigating and rectifying social biases. Kamboj et al. (2023) proposed a gender bias debiasing approach for the T5 model, which contextualized word embeddings. Their method involved adjusting embeddings to move towards the center of the gender polarity distribution. Another study (Dwivedi et al., 2022) pointed out that PE and ICL can be used for regulating LLM outputs, indicating that while LLMs may inherit biases, their manifestation can be controlled to some extent.

However, it's crucial to acknowledge that LLMs themselves can also exhibit societal biases. In a separate study, Fang et al. (2023) explored beyond word analysis to the document level, revealing significant disparities between AI-generated content by LLMs and that of reputable sources such as The New York Times and Reuters. Their findings indicated that LLM-generated content conveyed sentiments and toxicities toward various gender or race-related population groups and demonstrated discrimination against underrepresented communities. Our project will build upon these findings to further explore and address biases inherent in LLMs, paving the way for more equitable and fair textual communication.

This study will utilize the StereoSet dataset, which is designed to measure stereotypical bias in language models across gender, race, religion, and profession. It was developed by Nadeem et al. and published in 2021, which can be accessed on GitHub and the paper with code.

This research aims to investigate whether widely used Large Language Models (LLMs), such as ChatGPT, can "cleanse" stereotypes and biases. Traditional bias measurement methods often employ dictionary-based or qualitative approaches, which can be time-consuming and challenging to apply across different languages, temporal spans, and stereotype categories. Word embedding, a technique that represents words as vectors through deep neural networks, offers a solution by revealing semantic relationships based on their positions in space. Similar words are closer in this space, allowing word embedding to 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.

Building upon this methodology, our study will utilize word embedding to calculate biases. Specifically, given a corpus, a word2vec model will be trained to obtain vector representations for each word within the textual context. Distinct sets of words on different genders and ethnicities will be established (e.g., “she,” “her,” “women,” and “female” for females), and an average vector representing each gender or ethnicity will be computed by word embeddings. Subsequently, a Euclidean distance difference between these average vectors (e.g., males and females) and a word embedding of occupation (e.g., nurse) or adjective (e.g., intelligent) will be calculated. Finally, the average distance differences across various occupations or adjectives will serve as the bias score, with higher values indicating more pronounced biases.

To investigate the effectiveness of LLMs in correcting biases, our study will follow a 3x2x2x2 experimental design. Before experiments, gender bias scores and ethnic bias scores of the original texts will be calculated and recorded. 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, including 1, 2, and 3 times; "type of LLM" refers to the LLM product and version used, including GPT 3.5 and GPT 4.0; "LLM instruction" pertains to the commands given to the LLM for generation, including "Rewrite." and "Rewrite. Be aware of any potential bias."; and "type of bias" includes gender bias and ethnic bias. Therefore, our study is creative and innovative. It extends beyond solely examining differences in biases under singular conditions; we also explore the effects across different experimental setups, including multiple generations by the LLM and under passive ("Rewrite.") and proactive ("Rewrite. Be aware of any potential bias.") instructions, to assess their efficacy in "cleansing" biases.

To shed light on the effectiveness of current approaches, we aim to delve into the diverse policies and initiatives implemented by major players in the LLM landscape, including OpenAI (creator of ChatGPT), Llama (a project by Meta), Gemeni (developed by Google), and Copilots (Microsoft's project). By scrutinizing their methodologies and outcomes, we aim to discern the extent to which these efforts succeed in curbing biases inherent in the generated content. Furthermore, we aim to use our research results to examine the effectiveness of current LLMs in correcting biases, especially on gender and race. We hope this project can serve as the fruit of thoughts for policymakers and the public to see if LLM can be used to improve the embedded biases in the existing content and be the agent for social change in the future.

References

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

Dwivedi, S., Ghosh, S., & Dwivedi, S. (2023, December 14). Breaking the Bias: Gender Fairness in LLMs Using Prompt Engineering and In-Context Learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, *15*(4). https://doi.org/10.21659/rupkatha.v15n4.10

Dwivedi, S., Ghosh, S., & Dwivedi, S. (2023, December 14). Breaking the Bias: Gender Fairness in LLMs Using Prompt Engineering and In-Context Learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, *15*(4). https://doi.org/10.21659/rupkatha.v15n4.10

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

Katsarou, S., Rodríguez-Gálvez, B., & Shanahan, J. (2022, April 11). Measuring Gender Bias in Contextualized Embeddings. *AAAI Workshop on Artificial Intelligence With Biased or Scarce Data (AIBSD)*. https://doi.org/10.3390/cmsf2022003003

Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/2021.acl-long.416