Investigating and Mitigating Undesirable Biases in Large Language Models

Mahammed Kamruzzaman

University of South Florida
kamruzzaman1@usf.edu

Abstract

The rise of large language models (LLMs) has revolutionized natural language processing, offering immense capabilities across various applications. The widespread integration of these models into commonplace technology has brought to light deep concerns about the biases they encompass, which could serve to perpetuate negative preconceptions and social injustices. The scope of my research includes social biases, brand biases, the impact of personas on bias, and stereotypes in low-resource languages. My contributions aim to deepen our understanding of these biases and develop methodologies to mitigate them, enhancing the fairness and utility of LLMs across diverse global applications.

Introduction

Alongside the impressive new capabilities of recent language generation models such as ChatGPT, Llama, Mistral, and Gemini, these systems are increasingly involved in consequential decisions made in the real world. This includes job hiring and performance reviews, with tips for hiring managers appearing across the internet. Despite these advancements, LLMs continue to struggle with embedded biases, which raises questions regarding the ethical use of LLMs in real-life applications. Prior work focuses particularly on whether these AI systems produce specific stereotypes of underrepresented minorities. Furthermore, some researchers have extended these investigations to non-English languages, proposing new datasets aimed at measuring these biases more broadly. Moreover, beyond social stereotypes, LLMs exhibit biases of popularity, often favoring well-known items or ideas over lesser-known ones. The narrative shifts as recent investigations bring to light an underrepresentation of diverse cultural knowledge within LLMs, emphasizing a pronounced cultural bias. Addressing these social biases in LLMs is crucial for developing AI systems that are both fair and inclusive. The integration of AI in daily life necessitates addressing inherent biases within LLMs. My research addresses this by identifying biases in LLMs, exploring their consequences, and proposing interventions to reduce their impact.

Completed Studies

Social Bias in LLMs. Bias in NLP systems along the lines of gender and ethnicity has been widely studied, especially for specific stereotypes (e.g., *Asians are good at math*). I investigate bias along less-studied but still consequential, dimensions, such as age, beauty, beauty in profession, and institution, measuring subtler correlated decisions that LLMs make between social groups and unrelated positive and negative attributes. Although these subtler biases are understudied they follow people as much as gender and ethnicity do. I introduce a template-generated dataset of sentence completion tasks that asks the model to select the most appropriate attribute to complete an evaluative statement about a person described as a member of a specific social group. I report the correlations that I find for 4 LLMs (GPT-4, Llama-2, PaLM-2, Mistral) using Kendall's τ test. My findings indicate that all four models exhibit statistically significant biases for these less studied bias categories (Kamruzzaman, Shovon, and Kim 2024). Moreover, in an effort to extend the focus of bias research beyond English-centric datasets, I proposed ***BanStereoSet***, a dataset designed to evaluate 9 types of stereotypical social biases in multilingual LLMs for the Bangla language (Kamruzzaman et al. 2024a). This dataset not only serves as a crucial tool for measuring bias in multilingual LLMs but also facilitates the exploration of stereotypical bias, potentially guiding the development of more equitable language technologies in Bangladeshi contexts.

Brand Bias and Global vs. Local Preferences. LLMs not only exhibit social biases, but they also display other forms of bias. So, I also examine the biases exhibited by LLMs towards different brands, a significant concern given the widespread use of LLMs in affected use cases such as product recommendation and market analysis. Biased models may unfairly favoring established global brands while marginalizing local ones. Using a curated dataset across four brand categories (shoe, cloth, electronics, and beverage), I probe the behavior of LLMs in this space (Kamruzzaman, Nguyen, and Kim 2024). I find a consistent pattern of bias in this space—both in terms of disproportionately associating global brands with positive attributes and disproportionately recommending luxury gifts for individuals in high-income countries. I also find LLMs are subject to country-of-origin effects which may boost local brand preference in LLM out-

puts in specific contexts.

Mitigating Social Bias through Cognitive Frameworks. Dual process theory posits that human cognition arises via two systems. System 1, which is a quick, emotional, and intuitive process, which is subject to cognitive biases, and System 2, is a slow, onerous, and deliberate process. NLP researchers often compare zero-shot prompting in LLMs to System 1 reasoning and chain-of-thought (CoT) prompting to System 2. I investigate the relationship between bias, CoT prompting, debiasing prompt, and dual process theory in LLMs directly. I compare zero-shot CoT, debiasing, and a variety of dual process theory-based prompting strategies on two bias detection datasets spanning nine different social bias categories. I incorporate human and machine personas to determine whether the effects of dual process theory in LLMs exist independent of explicit persona models or are based on modeling human cognition. I find that a human persona, debiasing, System 2, and CoT prompting all tend to reduce social biases in LLMs, though the best combination of features depends on the exact model and bias category—resulting in up to a 19 percent drop in stereotypical judgments by an LLM (Kamruzzaman and Kim 2024b).

Personas and their Perceptions. As the deployment of LLMs expands, there is an increasing demand for personalized LLMs. One method to personalize and guide the outputs of these models is by assigning a persona—a role that describes the expected behavior of the LLM (e.g., a man, a woman). I investigate whether an LLM's understanding of social norms varies across assigned personas (Kamruzzaman et al. 2024b). Ideally, the perception of a social norm should remain consistent regardless of the persona, since acceptability of a social norm should be determined by the region the norm originates from, rather than by individual characteristics such as gender, body size, or race. In my research, I tested 36 distinct personas from 12 sociodemographic categories (e.g., age, gender, beauty) across four different LLMs. I find that LLMs' cultural norm interpretation varies based on the persona used and the norm interpretation also varies within a sociodemographic category (e.g., a fat person and a thin person as in physical appearance group) where an LLM with the more socially desirable persona (e.g., a thin person) interprets social norms more accurately than with the less socially desirable persona (e.g., a fat person). In my another research, I examine how the LLMs' perceptions of countries change when I assign nationality personas (e.g., an American person) to LLMs (Kamruzzaman and Kim 2024a). I assign 193 different nationality personas and find that all LLM-persona combinations tend to favor Western European nations, though nation-personas push LLM behaviors to focus more on and view more favorably the nation-persona's own region. Eastern European, Latin American, and African nations are viewed more negatively by different nationality personas.

Anticipated Future Work

I am currently working on a project where I investigate whether the emotional judgement of LLMs changes based on users' country-of-origin. Additionally, I am exploring how intersectional biases may affect the LLMs' emotional responses. For example, given a sentence that expresses a specific emotion (e.g., joy, sadness, etc.), I examine how the perception shifts based on age-gender or gender-nationality combinations, compared to single attributes like age or gender alone. I am also experimenting with techniques for mitigating social biases by using detailed CoT and self-evaluation together to see if I can reduce social biases on a larger scale.

Contributions and Collaborations

The research presented here was conducted in collaboration with my advisor and peers. I led all the studies, while my advisor provided guidance on the theoretical framework and methodological approaches, and my peers assisted with data collection.

Conclusion

My work contributes to a deeper understanding of the mechanisms through which biases manifest in LLMs and offers practical solutions for their mitigation. As LLMs continue to permeate various aspects of life, addressing these biases is critical to developing AI that is not only powerful but also equitable and respectful of global diversity.

References

Kamruzzaman, M.; and Kim, G. L. 2024a. Exploring Changes in Nation Perception with Nationality-Assigned Personas in LLMs. *arXiv preprint arXiv:2406.13993*. (Under review at NAACL 2025).

Kamruzzaman, M.; and Kim, G. L. 2024b. Prompting techniques for reducing social bias in llms through system 1 and system 2 cognitive processes. *arXiv preprint arXiv:2404.17218*. (Under review at AAAI 2025).

Kamruzzaman, M.; Monsur, A. A.; Das, S.; Hassan, E.; and Kim, G. L. 2024a. BanStereoSet: A Dataset to Measure Stereotypical Social Biases in LLMs for Bangla. *arXiv preprint arXiv:2409.11638*. (Under review at COLING 2025).

Kamruzzaman, M.; Nguyen, H.; Hassan, N.; and Kim, G. L. 2024b. "A Woman is More Culturally Knowledgeable than A Man?": The Effect of Personas on Cultural Norm Interpretation in LLMs. *arXiv preprint arXiv:2409.11636*. (Under review at COLING 2025).

Kamruzzaman, M.; Nguyen, H. M.; and Kim, G. L. 2024. "Global is Good, Local is Bad?": Understanding Brand Bias in LLMs. *In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. (Accepted, In Press).

Kamruzzaman, M.; Shovon, M.; and Kim, G. 2024. Investigating Subtler Biases in LLMs: Ageism, Beauty, Institutional, and Nationality Bias in Generative Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics ACL 2024*, 8940–8965. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics.