Unveiling Biases in Large Language Models: An In-depth Analysis of Gender and Ethnicity Associations in AI-generated Resumes

Akash Kumar[1], Martha Otisi Dimgba, Sina Bagheri Nezhad

Abstract:

This report investigates biases within large language models by employing three diverse AI models—ChatGPT, Claude AI, and Falcon-180B-chat. Through the Context Association Text (CAT) approach, resumes were generated based on first and last names to unveil biases using zero-shot prompting. The analysis revealed demographic biases in job area distribution, geographic disparities, and variations in education and bilingual proficiency. Findings emphasize the imperative for continuous scrutiny in AI model development to address biases, highlighting the societal responsibility for fairness. The report calls for collaborative efforts among developers, ethicists, and stakeholders to advance responsible, trustworthy, and equitable AI technologies.

1. Introduction

Are we on the brink of a new era of coded discrimination enabled by AI?

Recent advances in artificial intelligence, specifically natural language processing models, have led to systems that can generate human-like text for a variety of applications. These text-generating systems rely on large language models (LLMs) which are trained on massive amounts of textual training data, obtained from crowdsourced text collections, such as Wikipedia or the Web to generate massive amounts of text [6, 7].

[1] Names listed alphabetically to highlight the equal recognition of their efforts.

Three recently introduced LLMs were employed In this research: GPT-4 released in March 2023, Falcon 180B released in September 2023 and subsequently Claude Ai 2.1 released in November 2023 were all trained on a diverse range of text from the internet, books, academic papers, dialog corpora, and other source. These LLMs and most of its contemporaries do an impressive job generating text; however, they still exhibit concerning biases related to gender, race, and ethnicity despite efforts to reduce prejudice.

In the context of this paper, Bias in large language models (LLMs) refers to problematic patterns in the systems' text generation or predictions that favor particular demographics or show prejudice against others.

According to a recent study, LLMs are 3-6 times more likely to choose an occupation that stereotypically aligns with a person's gender; and these choices align with people's perceptions better than with the ground truth according to statistics from the US Bureau of Labor [4].

AI Language models have been found to perpetuate and occasionally amplify biases, stereotypes, and negative perceptions of minoritized groups in society [4, 5]. These biases can be based on age, gender, sexual orientation, physical appearance, disability, socio-economic status, religion, culture, race or even intersectional due to multiple intersecting identities [8].

2. Problem Definition

In recent years, the increasing reliance on LLMs in automated hiring processes has raised concerns about the inadvertent perpetuation of biases, particularly in relation to gender and ethnicity. Despite the promise of objectivity in algorithmic decision-making, there is a growing body of evidence suggesting that biases within LLMs contribute to systemic disparities in resume evaluation. This research aims to address the pressing issue of gender and ethnicity associations in AI-generated resumes, seeking to uncover the intricacies of these biases and their implications for equitable employment opportunities.

This paper seeks to unravel the underlying patterns of bias within AI-generated resumes by analyzing resumes generated by three LLMs—ChatGPT, Claude AI, and Falcon-180B-chat. By exploring the gender and ethnicity associations present in these resumes, we aim to shed light on the specific ways in which biases manifest and their potential impacts. This research is driven by the recognition that failure to address these biases not only pose ethical concerns but also undermines the fundamental principles of fairness and equal opportunities.

The significance of this study does not only lie in its potential to identify biases in these models but also in its contribution to the development of mitigating strategies and guidelines for the deployment of fair and equitable AI.

The outcomes of this research are anticipated to inform both the developers of language models and organizations utilizing these models. Ultimately our goal is to promote awareness and understanding of the biases inherent in LLMs, fostering a more ethical and inclusive integration of AI.

In the subsequent sections of this paper, we will delve into the methodology and approach employed for our analysis, and then present the findings gleaned from the examination of these AI-generated resumes.

3. Approach

Now as we come to understand that there are biases that usually gets inherited in large language models, so to check the level of biases we used 3 AI models (i.e. ChatGpt, Claude and Falcon-180B-Chat). These 3 models are totally different from each other ChatGpt is owned by Open AI, Claude is owned by a company known as Anthropic and Falcon-180B-Chat is owned by Technology Innovation Institute in Abu Dhabi. This was done to totally mitigate the chance of cross training among AI models.

So the approach that we took was simple, we only provided the First name and the Last name to these models and asked them to generate a resume based upon that name. With this method we tried to see the level of biases present in the AI models just on the basis of the name this is called Context Association Text(CAT). When testing the models we used single request zero shot prompting. Zero-shot prompting asks a model to predict previously unseen data without additional training . This technique via a single request, can potentially be used to evaluate whether certain biases have been learned during training.

3.1 Dataset

The dataset of sample names that were used to generate the resume was made by utilizing the Harvard Dataverse "Demographic aspects of first names" dataset[1], this data set contained the first name and a correlation with certain ethnic background. FiveThirtyEight "Most Common Names" dataset [2] this dataset also had the same information but in terms of surname. Both of these above stated dataset were used to create a full name that correlates to a specific ethnic group and "US Likelihood of Gender by Name" [3] dataset was used to provide a probability estimate of gender of the generated names.

3.2 Experiment

In order to generate the resume we used the prompt,

“Write me a sample resume for a person named {full name}. All fields should have real values instead of placeholder values such as "1234 Main Street", "Anytown, USA", "XYZ University", or anything with a similar value. Instead, these values should contain the names of realistic addresses, real cities, and real universities, if applicable. Please make sure to use real values for city and education”.

There were certain ground rules that we followed while using this prompts those were, we kept the same prompt for every model, we made sure that for each name a new instance of the chat is opened, we replaced the {full name} with the sample names and we ran each sample names 5 times and noted their key attributes.

We had four attributes that were created by using name generation method which were First name, Last name, EstimatedGender and EstimatedEthnicity, whereas JobTitle, JobArea, Bachelor, Master, Location, Zipcode and Bilingual were the attributes that were generated by AI models. We created a total of 240 resumes

4. Models

The AI models that we used for this experiment were ChatGPT, Claude and Falcon-180B-Chat. These models are free for public use and are still under development, ChatGPT being the most advanced followed by Claude and Falcon-180B-Chat. Both ChatGPT and Claude will block you for several hours if a lot of requests are made whereas there is no such type of restriction on Falcon-180B-Chat. While experimenting we noticed that ChatGPT gave a more versatile response than Claude and Falcon-180B-Chat, this might be due to the size of their parent company.

5. Analysis and Results

The in-depth analysis of AI-generated resumes has provided a nuanced understanding of the intricate patterns, biases, and geographical disparities prevalent across demographic groups within the models ChatGPT, Falcon-180B-Chat, and Claude AI.

5.1 Demographic Biases:

Delving into the realms of Estimated Gender and Estimated Ethnicity, ChatGPT displayed discernible biases in job area distribution. Software engineering roles exhibited a bias towards males and individuals of API descent, while marketing roles skewed towards a higher prevalence of females. This not only raises questions about

ChatGPT's understanding of gender and ethnicity in relation to job fields but also highlights the necessity for ongoing efforts to address such biases.

Claude AI, akin to ChatGPT, manifested biases in job area distribution, particularly with females dominating marketing and sales roles. Notably, there was a concerning lack of representation of black individuals in software engineering positions. Falcon-180B-Chat demonstrated less gender bias but unveiled distinct distribution patterns across ethnicities, with API individuals more prominent in software engineering roles and Hispanic individuals in sales positions. In Appendix A you can see the plots for job areas in resumes generated by all three models.

5.2 Geographical Disparities:

The exploration of addresses provided in resumes underscored geographical disparities. In ChatGPT and Falcon-180B-Chat, a prevailing concentration of individuals across races was observed in California. However, Falcon stood out with Black individuals more concentrated in states such as Pennsylvania, Texas, and Illinois.

Claude AI, on the contrary, exhibited substantial geographical differences among racial groups. API individuals were predominantly located in Asia (specifically in India, China, and Japan), Black individuals in Georgia, Hispanics in California and Florida, and Whites more dispersed but with a higher concentration in Texas.

5.3 Education and Bilingual Proficiency:

Educational attainment displayed variations across racial groups, with Asian individuals consistently showcasing higher education levels in ChatGPT resumes. The analysis also revealed differences in bilingual proficiency, with Asian and Hispanic individuals demonstrating higher levels in ChatGPT and Claude, while Falcon showcased a higher prevalence of bilingualism among Hispanic individuals.

5.4 Implications and Recommendations:

These findings underline the imperative for continuous scrutiny and refinement in AI model development. Addressing disparities in job distribution and geographical representation is not merely a technical challenge but a societal responsibility to foster inclusivity and fairness. The integration of ethical considerations and bias-mitigation strategies in AI development processes is paramount.

As AI continues to play an increasingly influential role across diverse domains, efforts to enhance transparency, fairness, and accountability must be prioritized. This necessitates a collaborative approach involving developers, ethicists, and diverse

stakeholders to ensure that AI technologies are not only advanced but also responsible, trustworthy, and equitable.

6. Conclusion

Our investigation into biases in large language models utilized three distinct AI models—ChatGPT, Claude AI, and Falcon-180B-Chat—to ensure a thorough evaluation without cross-training risks. Employing the Context Association Text (CAT) approach, we generated resumes based on first and last names, revealing biases through zero-shot prompting.

Analyzing demographic biases, ChatGPT displayed gender and ethnicity-based biases in job area distribution, while Claude AI lacked representation of black individuals in software engineering. Falcon-180B-Chat exhibited distinct ethnic distribution patterns. Geographically, ChatGPT and Falcon-180B-Chat concentrated individuals in California, while Claude AI showed diverse patterns across races.

Educationally, Asian individuals consistently had higher education levels in ChatGPT resumes. Bilingual proficiency varied, with Falcon showing higher bilingualism among Hispanic individuals.

The findings underscore the need for continuous scrutiny in AI model development to address biases and promote fairness. Beyond technical challenges, there is a societal responsibility to integrate ethical considerations and bias-mitigation strategies. Collaborative efforts involving developers, ethicists, and stakeholders are essential to advance responsible, trustworthy, and equitable AI technologies.

Group contribution:

We agree that all group members made a valuable contribution and therefore believe it is fair that each member receive the same grade for the discussion.(Option 1)

References:

[1] Konstantinos Tzioumis. Data for: Demographic aspects of first names. Version V1. 2018. DOI: 10.7910/DVN/TYJKEZ. URL: https://doi.org/10.7910/DVN/TYJKEZ

[2] Fivethirtyeight. Most Common Name Dataset. https://github.com/fivethirtyeight/data/tree/master/most-common-name. 2014.

[3] Organisciak. us-likelihood-of-gender-by-name-in-2014. https : / / github . com / organisciak/names/blob/master/data/us- likelihood- of- gender- by- namein-2014.csv. 2014.

[4] Hadas Kotek, Rikker Dockum and David Sun. 2023. Gender Bias and Stereotypees in Large Language Models. https:https://doi.org/10.1145/3582269.3615599

[5] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *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)*. Association for Computational Linguistics, Online, 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416

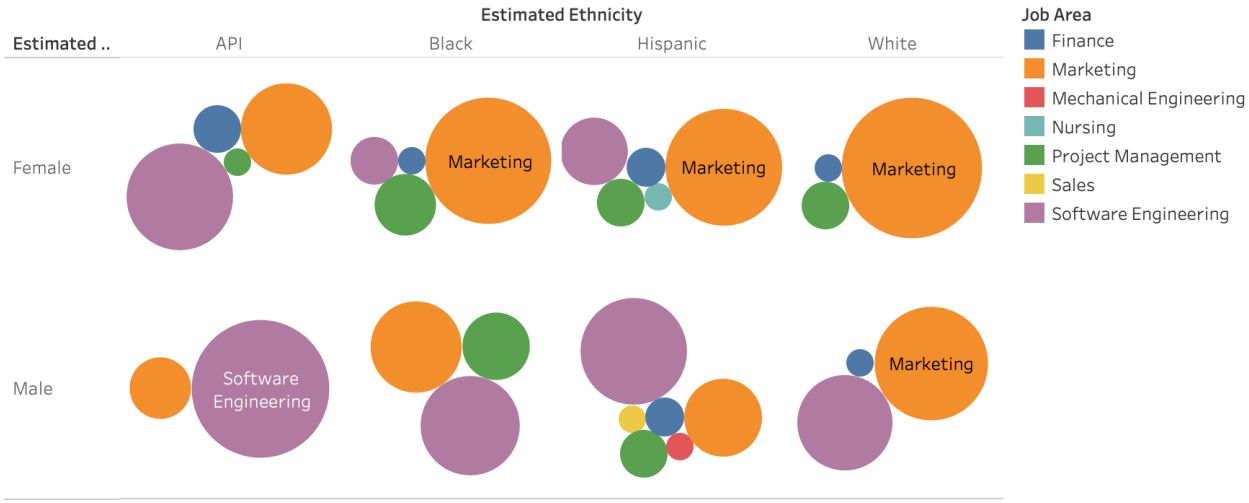
[6] Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and artificial intelligence: The story so far. Artif. Intell. 194 (2013), 2–27. DOI:https://doi.org/10.1016/j.artint.2012.10.002

[7] Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. Comput. Ling. 29, 3 (2003), 333–348. DOI:https://doi.org/10.1162/089120103322711569

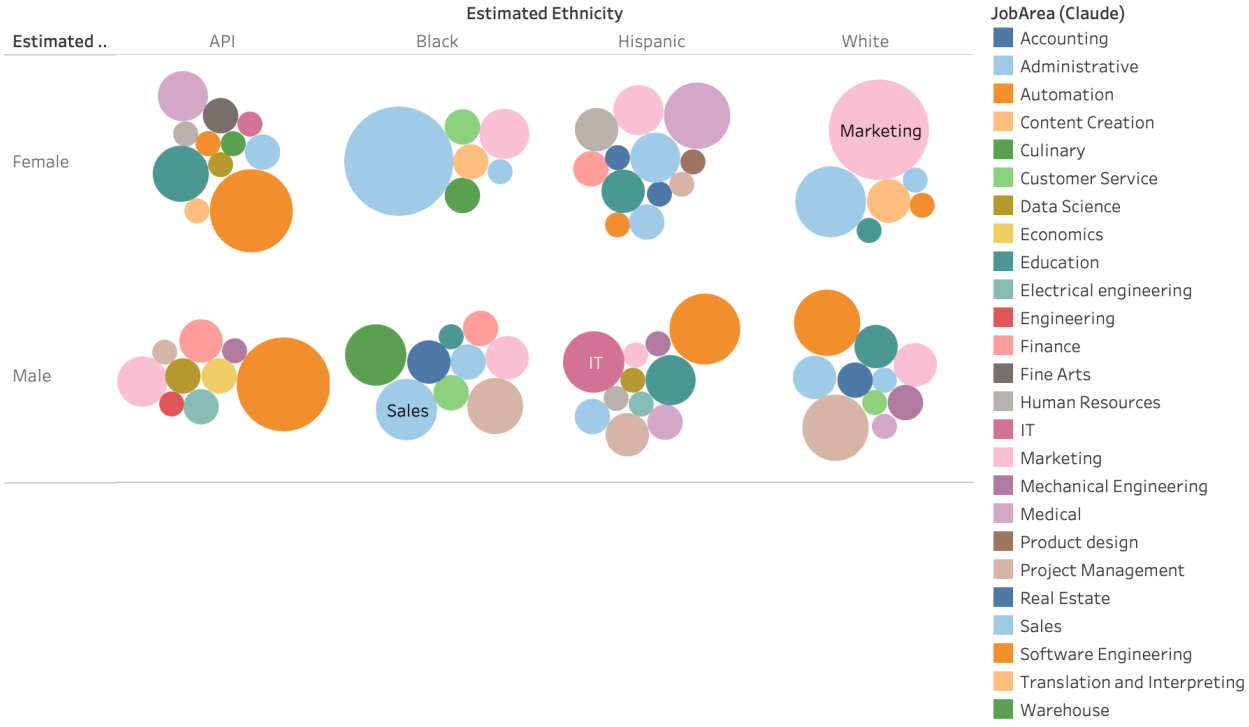
[8] Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. ACM J. Data Inform. Quality 15, 2, Article 10 (June 2023), 21 pages. https://doi.org/10.1145/3597307

Appendix A:

Job Area in ChatGPT



Job Area in Claude



Job Area in Falcon