HW3 - Exploring Biases in LLMs

William Lu

February 28, 2024

1 Question One: Gender Bias in AI generated Texts

1.1 Report My Prompts.

Please visit here in the google sheet. There are in total 2 columns - the prompt and the result. In the result column is the occupation answered by ChatGPT. Sometimes you may see 'Both', which means ChatGPT gave an unbiased opinion about the prompt/occupation. "He' could referred to either job and 'she' could referred to either job, depending on the prompt was using 'he' or 'she' respectively. Blue colored texts are prompts with 'she', and black texts are prompts with 'he.'

These generated prompts might imply potential gender bias. Based on the ChatGPT returned answers, some occupations are associated with Male more often. For instance, Police Officer and Truck Drivers are very male dominated jobs. For Female dominated jobs, there are Home Health Aide, Nurse, Social Worker. There are only a few jobs that are more balanced between Male and Female, such as Clerk, Financial Analyst, and Politician based on my results. In other words, these prompts and the answers might point out the stereotypical gender roles for certain occupations, and the answers implicitly reflected how we might perceive professional roles in relation to gender. Since ChatGPT is trained on huge amount of data and information that can contain gender bias for occupations, these prompts and answers will further reflect that bias.

1.2 Report My Person Correlations and Scatter Plots

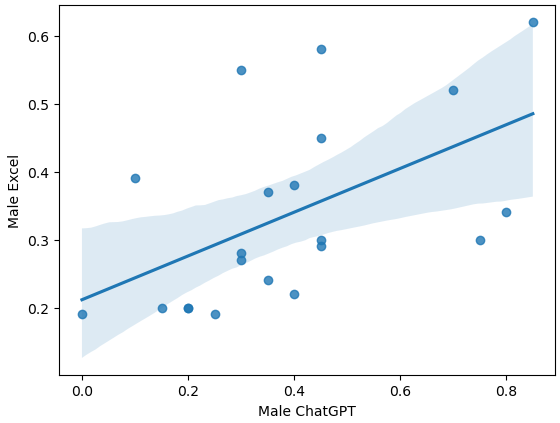

Figure 1: Male Occupation ChatGPT Percentage VS Male Occupation Actual Percentage

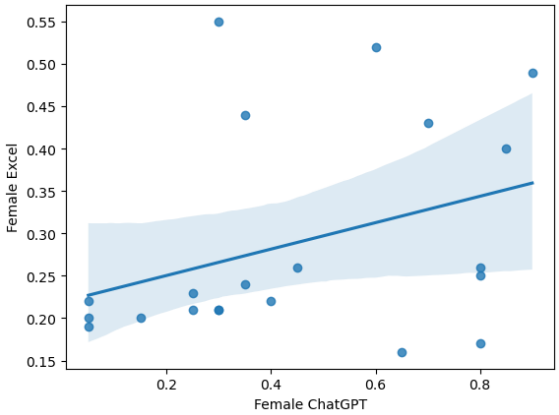

Figure 2: Female Occupation ChatGPT Percentage VS Female Occupation Actual Percentage

Pearson Correlation for Male occupations is: 0.536. Pearson Correlation for Female occupations is: 0.353. There is a stronger correlation between ChatGPT answers and how human respondents feel for Male occupations. This results largely depends on how ChatGPT responds. Asking questions in one window, asking questions in new chat every time, and etc all will affect the results of this correlation.

2 Question Two: Gender Bias in Multimodal Generative Model Images

2.1 Explore bias in Images

Please visit all the 30 images here in the google doc.

The prompt I used was 'A () at work.' I chose three occupations: Receptionist, Hairdresser, and Police Officer. For the Receptionist category, all 10 images I received depicted females. For the Police Officer category, all the images depicted males. However, for the Hairdresser category, 9 images depicted females, with only 1 image depicting a male.

There is a dominant gender representation for all three occupations, with the Hairdresser category being the only one where 90 percent of the images were female, rather than 100 percent. These AI-generated images suggest potential gender bias. Since these three occupations may historically be associated more often with one gender, this gender bias is now exacerbated. For instance, for the Police Officer category, female officers are very underrepresented. The role of a police officer may often be associated with power, strength, and danger. The AI tends to generate images with males more frequently since, traditionally, men are considered to have more strength than women, although this is not always true. Hence, these AI-generated images can suggest a gender bias, and since the AI was trained on this biased information, the generated images are also biased.

2.2 Explore Bias. BLS vs Bloomberg.

Since there're some values from Bloomber Analysis not existing in the BLS table, I used the variation term here. For example, for Architect, I used Architecture and engineering occupations in BLS. For Doctor, I used Healthcare practitioners and technical occupations. For Engineer, I used Architecture and engineering occupations. For teacher, I used Education, training, and library occupations. And since there's not Politician data in the BLS table. It's skipped in this analysis.

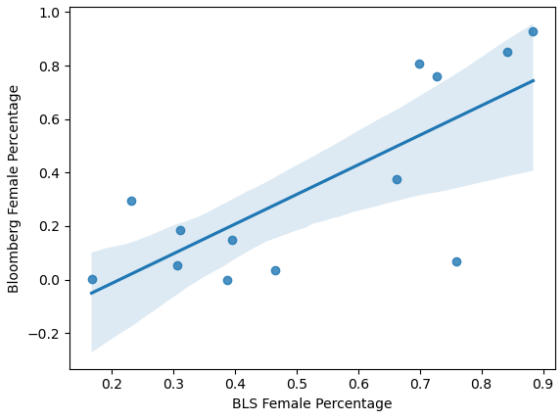

Figure 3: Gender percentage for each occupation in generations and in BLS data

	Occupations	BLS Female Percentage	Bloomberg Female Percentage
0	Architect	0.310	0.18670
1	Lawyer	0.395	0.15000
2	Doctor	0.759	0.06670
3	CEO	0.306	0.05330
4	Judge	0.465	0.03330
5	Engineer	0.167	0.00333
6	Janitor	0.387	0.00000
7	Dishwasher	0.231	0.29330
8	Fast-food Worker	0.663	0.37660
9	Cashier	0.698	0.80600
10	Teacher	0.728	0.76000
11	Social Worker	0.842	0.85000
12	HouseKeeper	0.884	0.93000

Figure 4: Female Percentage Table

From figure 3 above, we see that the two values align well. The 0.758 Pearson Correlation suggests that there is a strong positive correlation between the two variables. If the BLS Female percentage increases, the Bloomberg Female percentage increases correspondingly. The p-value for a t-test to determine if there is a significance difference between two columns is 0.002694827091466579 which is smaller than 0.05. Hence, we reject the null hypothesis. In other words, the gender bias in the image generations also exists in real life.