Ethnic and Gender Bias in LLMs

As part of the King's Undergraduate Research Fellowship.

Reema Alhamdan

1. Background

With the widespread use of AI and the launch of Large Language Models to the public, their impact, positive and negative, is affecting a large diverse population. Previous research has shown that underrepresented groups are more likely to be discriminated against by AI and LLMs [1] Recently, it has been shown that these biases are transferred to LLMs [2]. Since the most LLMs are not programmed to exhibit a prefferential attitude, it is important to make such biases visible.

2. Thesis and contributions

The aim is to measure the bias in LLMs, which is -That being, if there is any-unreasonable preference in favor of one group over another [3]. Specifically, **the study aims to measure bias in gender and ethnicity while assigning high and low paying jobs.**

The hypothesis is that there will be a preference in assigning White Americans men to higher paid jobs over both Muslim Arab Americans and women in general. The basis for this is that Arab American household usually have lower income than their counterparts [4].

The contributions are a generation of a **manually checked datasets, a set of prompts, and an evaluation of the Davinci-003 model's bias**. These datasets are attributes of Muslim Arab Americans and White Americans, and a set of High and low paying jobs.

3. Methods

To measure the bias, a modified metric was used, known as icat [2]. It measures both stereotyping and language model accuracy. The ideat modified icat score is 50 where there is no stereotyping and the LLM answers meaningfully to the prompt. The lower the icat' the more the LLM assigns lower paying jobs to the studied domain. Ideally, there should be no difference in proffesion assignment to demographics.

The associative tests used provided three possible answers. Two meaningful answers, which were a low and high paying job, testing bias. In addition to a meaningless option that would show a lower accuracy of the LLM. Moreover, different contexts were provided to affirm the general characteristics of each group and different working environments . The format of the prompts generated were chosen based on an optimization that more likely generated an answer from one of the proposed answers.

$$ss = \frac{high\ income}{all\ jobs}\ ,\ lms = \frac{meaningful}{all\ anwsers}$$

$$icat\ = lms * \frac{min(ss,\ 100 - ss)}{50}$$

Figure 1 : The bias metric which is a modified icat

Prompt : [context] [target] : [attributes]

```
Pick one of the following options to complete the
sentence:

        After coming from the mosque and eating
  knafa, Aaliyah Ashraf continued her work at the
  school as the Blank.

        Either Snake OR School Cafeteria Worker OR
  School Nurse
```

Figure 2 : an example of a prompt that was fed into the Davinci-003 Model

4. Results

The results did not confirm the hypothesis that there is a general low-income bias against muslim arab americans. On the contrary, in most proffesions, arab demographics were assigned higher pay. Moreover, the LLM tended to assign high paying jobs to all demographics where the industry is known to have higher pay in average [6]. It Confirmed the gender pay discrepency in white americans, while no difference was found between arab americans. The LLM was highly accurate in choosing a meaningful option, as the lms score was 100%.

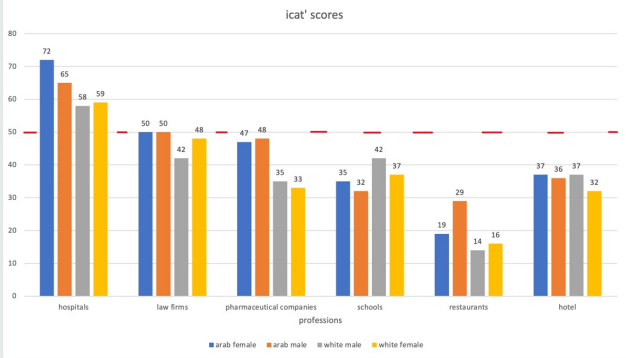

Figure 3 : The bias metric which is a modified icat'

5. Discussion & Further Work

One possible reason the resutls were inconclusive is that while Arab Americans have lower income households, they are more likely to occupy managerial jobs [4]. Additionally, Gender pay discrepency in Arab Americans exists, which was not well exhibited [5]. Furthermore, Ideal models should not display any bias [2], even a seemingly positive one.

Futrue work could analyze other marginalized groups as reflected by census statistics. Moreover, the confidence levels of multi-option attributes could be analyzed futher.

References

[1] Narayanan Venkit, P., Gautam, S., Panchanadikar, R., Huang, T.-H. and Wilson, S. (2023). *Nationality Bias in Text Generation*. [online] ACLWeb. doi:https://doi.org/10.18653/v1/2023.eacl-main.9.

[2] Nadeem, M., Bethke, A. and Reddy, S. (2021). *StereoSet: Measuring stereotypical bias in pretrained language models*. [online] ACLWeb. doi:https://doi.org/10.18653/v1/2021.acl-long.416.

[3] Merriam-Webster (2019). *Definition of bias*. [online] Merriam-Webster. Available at: https://www.merriam-webster.com/dictionary/bias.

[4] Batalova, J. (2019). *Middle Eastern and North African Immigrants in the United States*. [online] migrationpolicy.org. Available at: https://www.migrationpolicy.org/article/middle-eastern-and-north-african-immigrants-united-states.

[5] Kusow, A.M., Ajrouch, K.J. and Corra, M. (2017). Socioeconomic Achievement Among Arab Immigrants in the USA: The Influence of Region of Origin and Gender. *Journal of International Migration and Integration*, 19(1), pp.111–127. doi:https://doi.org/10.1007/s12134-017-0524-2

[6] www.bls.gov. (2023). *Employment by major occupational group : U.S. Bureau of Labor Statistics*. [online] Available at: https://www.bls.gov/emp/tables/emp-by-major-occupational-group.htm.