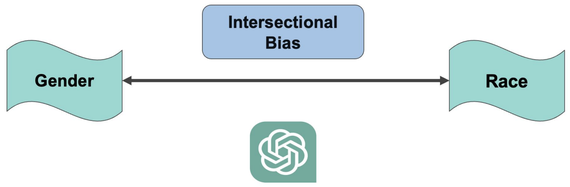
Fair GPT: Examine the Intersectional Bias in GPT-3.5

Hao Lin, Aishik Deb, Mainak Adak, Yuqing Wang

Stony Brook University

Introduction



Background

There still lacks assessment of intersectional biases between gender and race related to occupational associations for GPT-3.5, as the most state-of-the-art and also most accessible (because it is free) large language model. This study builds on this line of study on the social biases in generative language models, and contributes to provide an analysis of **occupational biases presented in GPT-3.5 for gender intersected with race using story generation approach**.

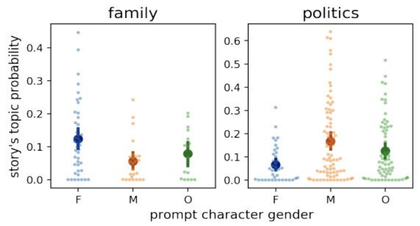


Figure by Li Lucy, David Bamman [6]

Figure by Abid A, Farooqi M, Zou J. [4]

Data and Methods

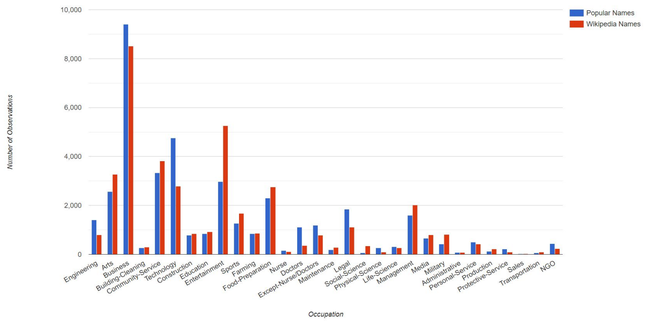
Data

Two base datasets for the purpose of extracting person names which are later used for story generation using name-based prompts

- Popular Names Dataset: NYC + Census, Top 50 names
- Wikipedia Names Dataset: Wikipedia ~ 22GB, Top 50 names

Name-based prompts. "write a story about [NAME] in the United States, do not exceed 100 words" ; a tall dataset having 80,000 rows and 36 features with occupation and demographics.

Occupation classification. 30 distinct occupational categories + GPT-3.5 to label data



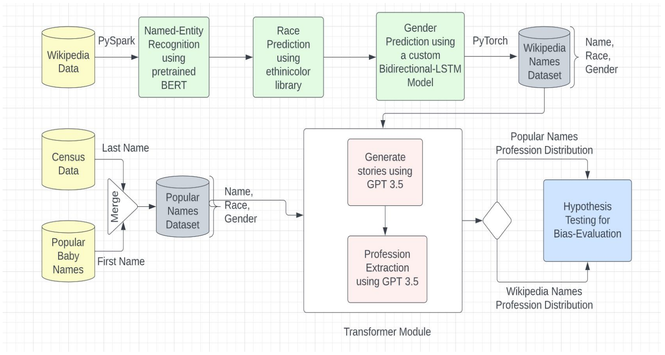
Methods

Name Entity Recognition: Bert-base-NER + pySpark

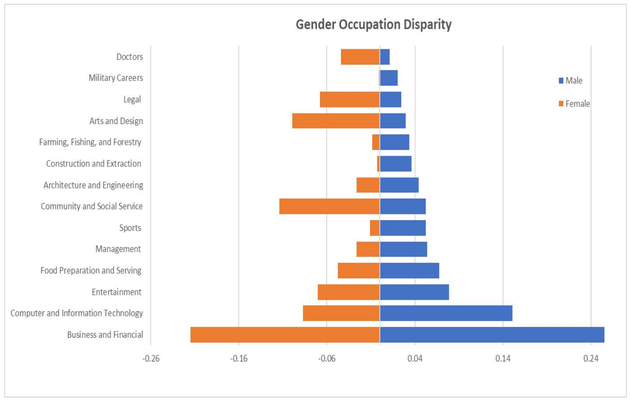
Gender Prediction Model: a bidirectional LSTM model

Hypothesis Testing: chi-square test + t-test (0.05)

- Gender differences: women with men
- Racial differences: White with other racial minorities
- Intersectional bias: White women with women of color & White men with men of color.



Results



Intersectional Bias	Popular Names Dataset	Wikipedia Names Dataset
	Jobs	Jobs
Female (Reference Group - White Female)		
Black	Food Preparation and Serving Construction and Extraction Sports	Community and Social Service Sports Arts and Design
Hispanic	Healthcare other than nurse and doctors Legal Community and Social Service	Food Preparation and Serving Construction and Extraction Building and Grounds Cleaning
Asian	Business and Financial Computer and Information Technology Arts and Design	Community and Social Service Computer and Information Technology Arts and Design
Male (Reference Group - White Male)		
Black	Sports Entertainment Military Careers	Military Careers Social Science Entertainment
Hispanic	Food Preparation and Serving Construction and Extraction Sports	Business and Financial Legal Food Preparation and Serving
Asian	Computer and Information Technology Business and Financial Architecture and Engineering	Computer and Information Technology, Education, Training, and Library Media and Communication

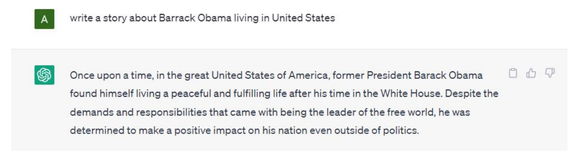
Racial biases in occupations

- **White and Black.** Results from two-sample t-test shows that GPT-3.5 has learnt the stereotypical association between sports and Black people.
- **White and Asian.** Compared to White, occupations that are dominant by Asian are less diverse; among those dominant by Asian many are stereotypically associated with Asian such as Engineering, technology, doctors, and food occupation.
- **White and Hispanics.** Compared to White, occupations that are dominated by Hispanics are less diverse; lower-skilled professions such as building and grounds cleaning are more likely to be associated with Hispanics than White.

Limitation

- Although names are often used as a proxy for demographics, they fall short in accurately capturing the complexity and diversity present within a population, and fails to account for the ambiguity of certain names.
- We only use 50 names for each social group and generate 100 stories for each name. Our results will be more reliable if we have more data points for more names.

Conclusion



- Using name-based prompts, our study identifies gender, racial and intersectional biases associated with occupations in GPT-3.5.
- This study also stresses the importance of looking at intersectionality, which helps us understand unique positionality and social experience of individuals in real life, and thus mitigate such bias and disparity.
- Since GPT-3.5 is already trained on Wikipedia text, the stories generated by it are about real people who are documented by wikipedia.

References

[1] Törnberg P. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. arXiv preprint arXiv:2304.06588. 2023 Apr 13.

[2] O Kjell, K Kjell, HA Schwartz. AI-based Large Language Models are Ready to Transform Psychological Health Assessment. 2023.

[3] Walby S, Armstrong J, Strid S. Intersectionality: Multiple inequalities in social theory. Sociology. 2012 Apr;46(2):224-40.

[4] Abid A, Farooqi M, Zou J. Persistent anti-muslim bias in large language models. InProceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society 2021 Jul 21 (pp. 298-306).

[5] Kirk HR, Jun Y, Volpin F, Iqbal H, Benussi E, Dreyer F, Shtedritski A, Asano Y. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. Advances in neural information processing systems. 2021 Dec 6;34:2611-24

[6] Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Acknowledgement and Contact

Thanks to Dr. Andrew Schwartz and Peter Geiss for their feedback and suggestions.

If you have any questions, please feel free to reach out to us at hao.lin@stonybrook.edu, aideb@cs.stonybrook.edu, madak@cs.stonybrook.edu, yuqing.wang.1@stonybrook.edu