Variability of Racial/Ethnic Groups in LLM-generated Stories

WashU CSE 527A Project

Messi H.J. Lee
Washington University in St. Louis
hojunlee@wustl.edu

Patrick Farley
Washington University in St. Louis
pfarley@wustl.edu

Sangwook Suh
Washington University in St. Louis
sangwooksuh@wustl.edu

1 Paper Summary

Title	The Effect of Group Status on the Variability of Group Representations in LLM-generated Text
Venue	Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2023
Year	2023
url	Unpublished work

Table 1: Bibliographical information for Anonymous (2023).

Background It is a known concern that Large Language Models (LLMs) reproduce human-like bias in their training data. Past work has documented biases pertaining to gender (Lucy and Bamman, 2021; Kotek et al., 2023), sexual orientation (Felkner et al., 2023), religion (Abid et al., 2021), politics (Motoki et al., 2023; Rozado, 2023), and culture (Huang and Xiong, 2023) in LLMs, among many others. Many of the aforementioned studies focus on stereotyping, associations of social groups with stereotypic attributes.

This paper, accepted for poster presentation at the Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2023 and attached with the proposal, aims to compare the homogeneity of texts generated for racial/ethnic and gender groups. This bias is grounded on the study of a social psychological phenomenon known as the *out group homogeneity effect* which pertains to the perception of outgroup members as more homogeneous than members of the ingroup. This phenomenon has been documented across many social group identities (Quattrone and Jones, 1980; Linville et al., 1989; Park and Judd, 1990; Ackerman et al., 2006).

Summary of contributions This paper focuses on quantifying bias in ChatGPT-generated texts about groups at the intersection of race/ethnicity and gender. They had ChatGPT generate 30 word entries, ranging from stories to social media profiles, about an African/Asian/Hispanic/White American man/woman. They obtained 500 completions for each writing prompt, induced sentence embeddings for the generated text, and compared the cosine similarity of the embeddings to evaluate the homogeneity of the generated texts. The paper found that cosine similarity values of the socially subordinate racial/ethnic groups (i.e., Asian, African, and Hispanic Americans) were significantly greater than those of White Americans, suggesting that ChatGPT portrays socially subordinate racial/ethnic groups as more homogeneous than the socially dominant racial/ethnic group. Similarly, they found that ChatGPT portrays the socially subordinate gender group (i.e., women) as more homogeneous than the socially dominant gender group (i.e., men), but to a smaller extent.

Limitations and discussion The biggest strength of this paper is that it sheds light on a form of human-like bias other than stereotyping - that subordinate social groups tend to be represented in a more homogeneous way compared to the dominant social group. However, one limitation is that the paper does not investigate where the homogeneity is coming from. Furthermore, it only uses one measure of similarity - cosine similarity of sentence embeddings - to compare the homogeneity of texts. Hence, studying the thematic overlap of the generated text may shed light on where the homogeneity is coming from, help generalize their findings using some other measure of homogeneity, promote better understanding of the harms that this bias may cause, and connect this bias to the study of stereotyping in LLMs. Finally, the study used ChatGPT to collect data, and it would be interesting to see

if the bias generalizes to other LLMs, particularly those that are more suited to generate stories.

Why this paper? Bias in LLMs is an incredibly important issue to tackle both for ethical and practical reasons. While the majority of research on bias has been focused on LLMs reproducing prevalent stereotypes (Blodgett et al., 2020), this paper instead focuses on a different, yet important, bias. Homogeneous representations of racial/ethnic groups in LLM-generated text may relate to new type of stereotypes reproduced by LLMs that have not yet been documented, and it may lead to more bias in future LLM training, which would in turn lead to more bias, as LLMs have been shown to be vulnerable to runaway feedback loops where subsequent training/updates to the model reinforce pre-existing biases in the training data (Ensign et al., 2018).

Wider research context This paper is related to the reproduction of human-like bias embedded in the training data of LLMs. The paper does not effectively mitigate or solve the identified bias and simply focuses on documenting a bias in LLMs that has previously been unexplored in the literature. This method of evaluating bias has promise in evaluating the bias within LLMs other than ChatGPT. Why LLMs reproduce this form of bias is worthy of further investigation as it may further promote research into mitigation strategies to help correct it.

2 Project Description

Goal The work introduced above has demonstrated that ChatGPT portrays socially subordinate social groups as more homogeneous than the socially dominant group using similarity measurements of sentence embeddings (Anonymous, 2023). However, it remains unseen whether the homogeneity extends to other measures of similarity (e.g., word overlap, thematic overlap). In this paper, we attempt to test whether texts written about racial/ethnic minority groups in the United States tend to share more thematic overlap compared to the racial/ethnic majority group (i.e., White Americans). We expect to see that the topic distribution of stories written for White Americans to be more dispersed compared to those of racial/ethnic minority groups.

Task The task that we propose to perform as part of the project are as follows. Instruct state of the art LLMs to generate stories about four racial/ethnic groups (i.e., African, Asian, Hispanic, and White Americans). Using the generated stories, fit structural topic models using the `stm` package in R. Compare the topic proportions of the stories generated for each of the racial/ethnic groups.

Data We propose collecting data from a number of state of the art LLMs including and not limited to ChatGPT, GPT-4, in addition to commercial models that are tailored for the purpose of generating stories (e.g., NovelAI, Dreamlily, and KoboldAI). As we need a moderately large data set to reliably fit a structural topic model, we will collect 100 stories per racial/ethnic group. To control for the lengths of the texts generated, we will set a length constraint to the LLM completions.

Methods We will use the `stm` package in R (Roberts et al., 2023) to fit a structural topic model on the generated stories. First, the `searchK()` function will be used to identify the ideal number of topics to identify from our collected stories. Then, we will fit a structural topic model using the `stm()` function. We will incorporate the race/ethnicity variable as document-level metadata so that the model accounts for race/ethnicity when identifying latent topics. This will also allow us to compare which topics are more prevalent for individual racial/ethnic groups and allow us to identify topics and attributes that a group is commonly associated with.

Baselines Here, we are interested in the idea of *group fairness* which points to the idea that LLMs (and any other Artificial Intelligence systems) are expected to perform similarly across groups (see Buolamwini and Gebru, 2018). Hence, the comparisons of the proposed project will happen within our data set where we compare the topic distribution of one group with the other. In this case, we are comparing the topic distribution of stories written for African, Asian, and Hispanic Americans to that of White Americans.

Evaluation Comparing the variability of topic distributions in textual data remains a relatively unexplored topic in the literature. However, we propose a test statistic we could use to quantify variability of topic distributions. First, we propose performing statistical tests that are used to compare the variability of distributions (e.g., Levene's test, F-test of equality of variances). The resulting F-statistic of the two tests quantifies the ratio between

the variance of the two distributions (σ_1^2 and σ_2^2; see Equation 1). If F is significantly greater than 1, it indicates that the variability of group 1's distribution is significantly greater than that of group 2.

$$F = \frac{\sigma_1^2}{\sigma_2^2} \quad (1)$$

We have a number of measurements that we can compare. First, we could compare the distribution of theta-values where theta corresponds to the topic distribution of an individual document. Second, we could instead assign an individual document to a single topic where the theta value is largest and compare the distribution of the single topics. Finally, we could use the theta values to induce a vector representation of an individual document and compare the distribution of cosine similarity values.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models.

Joshua M. Ackerman, Jenessa R. Shapiro, Steven L. Neuberg, Douglas T. Kenrick, D. Vaughn Becker, Vladas Griskevicius, Jon K. Maner, and Mark Schaller. 2006. They all look the same to me (unless they're angry): from out-group homogeneity to out-group heterogeneity. *Psychological Science* 17(10):836–840. https://doi.org/10.1111/j.1467-9280.2006.01790.x.

Anonymous. 2023. The effect of group status on the variability of group representations in llm-generated text.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pages 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485.

Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, pages 77–91. ISSN: 2640-3498. https://proceedings.mlr.press/v81/buolamwini18a.html.

Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway Feedback Loops in Predictive Policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, pages 160–171. ISSN: 2640-3498. https://proceedings.mlr.press/v81/ensign18a.html.

Virginia K. Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models.

Yufei Huang and Deyi Xiong. 2023. CBBQ: A Chinese Bias Benchmark Dataset Curated with Human-AI Collaboration for Large Language Models. ArXiv:2306.16244 [cs]. https://doi.org/10.48550/arXiv.2306.16244.

Hadas Kotek, Rikker Dockum, and David Q. Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*. pages 12–24. ArXiv:2308.14921 [cs]. https://doi.org/10.1145/3582269.3615599.

P. W. Linville, G. W. Fischer, and P. Salovey. 1989. Perceived distributions of the characteristics of in-group and out-group members: empirical evidence and a computer simulation. *Journal of Personality and Social Psychology* 57(2):165–188. https://doi.org/10.1037//0022-3514.57.2.165.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In Nader Akoury, Faeze Brahman, Snigdha Chaturvedi, Elizabeth Clark, Mohit Iyyer, and Lara J. Martin, editors, *Proceedings of the Third Workshop on Narrative Understanding*. Association for Computational Linguistics, Virtual, pages 48–55. https://doi.org/10.18653/v1/2021.nuse-1.5.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: measuring chatgpt political bias. *Public Choice* https://doi.org/10.1007/s11127-023-01097-2.

Bernadette Park and Charles M. Judd. 1990. Measures and models of perceived group variability. *Journal of Personality and Social Psychology* 59(2):173–191. Place: US Publisher: American Psychological Association. https://doi.org/10.1037/0022-3514.59.2.173.

George A. Quattrone and Edward E. Jones. 1980. The perception of variability within in-groups and out-groups: Implications for the law of small numbers. *Journal of Personality and Social Psychology* 38(1):141–152. Place: US Publisher: American Psychological Association. https://doi.org/10.1037/0022-3514.38.1.141.

Margaret Roberts, Brandon Stewart, Dustin Tingley, and Kenneth Benoit. 2023. stm: Estimation of the Structural Topic Model. https://cran.r-project.org/web/packages/stm/index.html.

David Rozado. 2023. The Political Biases of ChatGPT. *Social Sciences* 12(3):148. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/socsci12030148.