Quantifying Homogeneity Bias in Large Language Models: Entropy as a Measure of Text Heterogeneity across Groups

WashU CSE 527A Final Project Report

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

Abstract

Bias is a major concern in Large Language Models (LLMs), as they can perpetuate existing stereotypes. A significant yet unexplored this bias is *homogeneity bias*, which refers to the tendency of LLMs to represent certain social groups as less diverse than others. In this study, we contribute to the literature by 1) employing text completion prompts to examine jobs that LLMs associate with various gender and racial/ethnic groups, and 2) proposing a novel measure of homogeneity in LLM-generated text, namely entropy. This measure quantifies bias by assessing the uncertainty inherent in the variable outcome (i.e., jobs, industry) associated with these groups. Our analysis reveals two key findings: 1) Jobs associated with females are more homogeneous than those associated with males; and 2) Jobs associated with Asian Americans are more homogeneous than those associated with other racial/ethnic groups. These findings not only provide clear evidence of homogeneity bias in LLMs but also establish a critical benchmark for future research aimed at evaluating and mitigating this form of bias. [1]

1 Introduction

Bias in Large Language Models (LLMs) is a critical issue. These models, influenced by the biases in their training data, can produce outcomes that not only reflect but potentially amplify societal stereotypes (e.g., Bender et al., 2021; Blodgett et al., 2020; Nadeem et al., 2021). This is particularly concerning given the increasing integration of LLMs into daily life, where biased outputs can reinforce harmful stereotypes. Much research has focused on identifying and mitigating gender and ethnic stereotypes in LLMs. However, an area that requires further exploration is *homogeneity bias*, which pertains to the diversity of the model's outputs, specifically regarding the portrayal of different social groups (Lee et al., 2023).

[1] https://github.com/Skyblade64/527A-project

There are a number of reasons this bias may be a threat to targeted groups. By portraying groups as less diverse than others, the model risks disseminating incorrect or oversimplified representations about these groups. This can also lead to perpetuating stereotypes where the oversimplified representations of the groups may rely on existing societal stereotypes embedded in the training data. This can also have an influence on social discourse given the role that LLMs play in shaping both online and offline discourse. Homogeneous representations of groups may limit the scope of public discourse to a narrow range of perspectives and experiences, undermining efforts towards diversity and inclusion. Ultimately, homogenization of social groups in LLM outputs raises concerns about the fairness and equity of LLMs.

In this paper, we introduce a novel approach to assess homogeneity bias. First, we utilize text completion prompts to systematically examine the types of jobs that LLMs associate with various gender and racial/ethnic groups. This approach is more systematic as it limits the scope of the text generated by the LLM, allowing us to specifically quantify the heterogeneity of the completions. The method provides a distinct way to assess homogeneity bias, concentrating on a targeted aspect of the model's output. Secondly, we adopt entropy as a metric for analyzing this data, moving away from the commonly used cosine similarity. Entropy, as measure of the uncertainty inherent in a variable's outcomes, offers a direct assessment of the diversity of these completions.

Our findings reveal that GPT-4 tends to represent women and Asian Americans more homogeneously compared to men and other racial/ethnic groups. This pattern is less marked but still present in the context of industry (e.g., Healthcare, Technology and IT, among many others), highlighting a tendency in the model towards less varied portrayals for certain groups. These insights are vital, as they

not only point out the existence of homogeneity bias in LLMs, but also highlights the necessity for strategies that encourage more diverse and equitable representations in LLM-generated content.

2 Related Work

To the best of our knowledge, Lee et al. (2023) was the first to document homogeneity bias in LLM-generated text. Lee et al. (2023) designed writing prompts asking ChatGPT (`gpt-3.5-turbo`) to generate 30-word texts about eight groups at the intersection of two gender group identities and four racial/ethnic group identities. They encoded the generated text into sentence embeddings using models like Sentence-BERT (Reimers and Gurevych, 2019) and compared the pairwise cosine similarity values between sentence embeddings of the intersectional groups. They found that socially subordinate groups (i.e., women and racial/ethnic minorities) were portrayed as more homogeneous than their socially dominant counterparts and speculated that representation and stereotypicality of subordinate groups in the training data of LLMs may be the cause of this bias.

However, there are a number of limitations to this way of measuring text homogeneity. First, by encoding LLM-generated text into sentence embeddings using pre-trained language models like Sentence-BERT, they risk introducing an additional layer of bias (see Kurita et al., 2019), thereby making it hard to distinguish whether the bias is coming from the model that generated the text or the model that encoded it. Furthermore, sentence embeddings lack direct semantic interpretability. The dimensions of contextual embeddings do not map onto distinct, understandable semantic features, making it hard to interpret why two sentences are considered more similar than others. Hence, the study of homogeneity bias in LLMs could benefit from an assessment that does not rely on other language models that could potentially influence bias assessment and/or negatively impact interpretability.

3 Approach

To account for the limitations identified in previous work, we made two changes. First, instead of having an LLM generate longer forms of text, we had the model complete prompts with job titles. This allowed us to directly assess the homogeneity of jobs associated with gender and racial/ethnic groups without having to encode them into sentence embeddings. Second, to quantify the homogeneity of jobs associated with groups, we used entropy that directly quantifies variability or dispersion of the generated texts.

4 Experiment

In this experiment, we first design writing prompts to assess the jobs that GPT-4 associates with two gender groups (i.e., male and female) and four racial/ethnic groups in the US (i.e., African, Asian, Hispanic, and White Americans). We select names to signal gender and race/ethnicity in these prompts using the Name-Trait Perceptions data set (Elder and Hayes, 2023). Using the OpenAI API, we have GPT-4 complete the designed writing prompt. To determine if the groups are associated with broader industry categories (e.g., Healthcare, Technology and IT), we have GPT-4 classify the generated job titles into 20 industry categories. Then, we derive and compare the bootstrap confidence intervals for the entropy measure of the raw job titles and the industry categories that the titles have been assigned to.

4.1 Name Selection

We first selected a collection of names to represent eight groups at the intersection of four racial/ethnic (i.e., African, Asian, Hispanic, and White) and two gender (i.e., Male and Female) social groups. From the Name-Trait Perceptions data set, a data set of names and a a range of group identities (i.e., gender and race/ethnicity) and attributes (e.g., Democrat/Republican, aggressive, hardworking; Elder and Hayes, 2023). We randomly sampled 15 first names for each intersectional group,[2] totaling 120 names.

4.2 Generating Responses

The names were then inserted into a prompt crafted to get job titles as responses from GPT-4, specifically `gpt-4-1106` (as of 18 Nov 2023). We obtained fifty completions for each of the names, totaling 6,000 completions.

To prevent the model from avoiding to answer the prompt, we gave the model the following system instruction: *You are a chatbot. Don't generate notes. If you cannot determine the answer, guess. Answer with the name of a job and nothing else.* Then, the following writing prompt was given:

[2]The Name-Traits Perceptions data set only contained 37 names that were perceived as Asian, 16 of which were perceived as names of Asian men.

"Fill in the blank to the best of your ability. Guess if needed: The job of [name] is: _____."

4.3 Preprocessing

Prior to analyzing the data, we performed the following pre-processing steps:

1. Change to lowercase;
2. Trim whitespaces;
3. Remove punctuation;
4. Remove responses with five or more words.

Note that we removed responses with five or more words to filter out invalid responses. We were left with 5,972 completions after the pre-processing steps were completed. In Table 1, we show a randomly sampled completion for each of the intersectional groups.

Name	Race	Gender	Text
Kim	Asian	F	graphic designer
Wei	Asian	M	translator
Desiree	Black	F	librarian
Reginald	Black	M	gardener
Rosa	Hispanic	F	teacher
Ruben	Hispanic	M	electrician
Candice	White	F	dentist
Leon	White	M	translator

Table 1: Randomly sampled prompt and text completion for each intersectional group. Text column contains the text generated by GPT-4 after pre-processing.

4.4 Classification

Finally, we utilized the zero-shot classification capability of GPT-4 for categorizing the generated jobs into distinct industry categories (e.g., "Dentist", "Doctor" and "Surgeon" were all categorized as part of the "Healthcare" industry). Initially GPT-4 produced 18 industry categories. This list was complemented with two additional categories (i.e., "Sports Industry" and "Translation Services Industry") by prompting after manual inspection of the initial classification results. Furthermore, we iteratively prompted GPT-4 to include specific jobs into their matching industry categories as the majority of jobs was initially classified as "Other". For example, "Dentist" was initially classified as "Other" despite its apparent association with the "Healthcare Industry". In Table 2, we provide the industry categories associated with the eight jobs provided in Table 1. All 20 industry categories are listed in A.1.

Text	Industry
graphic designer	Creative Arts and Entertainment
translator	Translation Services
librarian	Education and Academic
gardener	Other
teacher	Education and Academic
electrician	Energy and Utilities
dentist	Healthcare
translator	Translation Services

Table 2: Job titles and the industry associated with the job titles for those jobs listed in Table 1.

4.5 Evaluation

In information theory, entropy is a measure of information content (MacKay, 2005). Given a discrete random variable X, with possible values $\{x_1, x_2...x_i\}$ and probability mass function $P(X)$, entropy is calculated using Equation 1.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i) \quad (1)$$

where $P(x_i)$ is the probability of the ith outcome, n is the number of possible outcomes, and b is the base of the logarithm calculations. b is most often set to 2 for binary systems such that entropy is measured in bits. Given the mathematical properties of the measure, entropy increases as the number of possible outcomes increases and as the distribution of the possible outcomes is closer to a uniform distribution.

Entropy stands out as an apt measure for quantifying heterogeneity in our data due to its ability to capture diversity in categorical data. Unlike variance, which primarily quantifies the spread of numerical data around a mean, entropy excels in evaluating the complexity and variability of non-numeric, categorical data like text. By measuring how evenly the information is distributed across different outcomes (in this case, job titles and industry categories), entropy reflects the degree of homogeneity/heterogeneity in LLM-generated text. This is particularly relevant in the context of homogeneity bias because we expect certain social groups to be depicted as more homogeneous than others. This could take two forms: 1) the group is associated with a narrower set of jobs, or 2) the distribution of jobs associated with a group is heavily skewed (e.g., a group is significantly more associated with a specific job compared to others).

To reliably compare the entropy measure derived from texts generated for different social groups and to account for the uncertainty in the measure, we performed bootstrapping. We sampled n texts generated for each of the intersectional groups, n being the number of texts for each group, with replacement, calculating entropy for both jobs and industry categories. We repeated this step 1,000 times to derive 95% Confidence Intervals (CIs) for each of the groups.

5 Results

Given our interest in the effect that gender and race/ethnicity have on the homogeneity of LLM-generated text, we interpret our results with reference to those laid out by Lee et al. (2023).

5.1 Gender Identity

Consistent with previous findings, we find that the both jobs and industry categories associated with the female gender are significantly more homogeneous compared to those of the male gender. Entropy value of male jobs (H_{male} = 5.26, 95% CI = [5.19, 5.34]) was significantly greater than that of female jobs (H_{female} = 4.68, 95% CI = [4.59, 4.76]), and the entropy value of male industry categories (H_{male} = 3.45, 95% CI = [3.41, 3.49]) was significantly greater than that of female industry categories (H_{female} = 3.36, 95% CI = [3.22, 3.41]; see Table 3).

	Job		Industry	
Gender	Entropy	95% CI	Entropy	95% CI
Male	5.26	[5.19, 5.34]	3.45	[3.41, 3.49]
Female	4.68	[4.59, 4.76]	3.36	[3.22, 3.41]

Table 3: The entropy of jobs and industry categories and their 95% confidence intervals of the two gender groups.

5.2 Race

Consistent with previous findings, we find that the both jobs and industry categories associated with Asian Americans are significantly more homogeneous compared to those of White Americans. Entropy value of White American jobs (H_{white} = 5.11, 95% CI = [5.01, 5.22]) was significantly greater than that of Asian American jobs (H_{asian} = 4.09, 95% CI = [3.96, 4.23]). Furthermore, entropy value of White American industry categories (H_{white} = 3.45, 95% CI = [3.39, 3.51]) was significantly greater than that of Asian American industry categories (H_{asian} = 2.91, 95% CI = [2.83, 3.00]; see Table 4).

	Job		Industry	
Race	Entropy	95% CI	Entropy	95% CI
Black	5.11	[5.02, 5.21]	3.41	[3.35, 3.47]
Asian	4.09	[3.96, 4.23]	2.91	[2.83, 3.00]
Hispanic	4.83	[4.73, 4.92]	3.57	[3.53, 3.62]
White	5.11	[5.01, 5.22]	3.45	[3.39, 3.51]

Table 4: The entropy of jobs and industry categories and their 95% confidence intervals of the four racial/ethnic groups.

Contrary to previous findings, however, we find that there is no significant difference between the jobs and industry categories associated with African and White Americans. Entropy value of White American jobs (H_{white} = 5.11, 95% CI = [5.01, 5.22]) was not significantly greater than that of African Americans (H_{black} = 5.11, 95% CI = [5.01, 5.22]). Furthermore, entropy value of White American industry categories (H_{white} = 3.45, 95% CI = [3.39, 3.51]) was not significantly greater than that of African American categories (H_{black} = 3.41, 95% CI = [3.35, 3.47]; see Table 4).

Finally, we find mixed patterns when comparing the jobs and industry categories associated with Hispanic and White Americans. Entropy value of White American jobs (H_{white} = 5.11, 95% CI = [5.01, 5.22]) was significantly greater than that of Hispanic Americans ($H_{hispanic}$ = 4.83, 95% CI = [4.73, 4.92]). However, entropy value of White American industry categories (H_{white} = 3.45, 95% CI = [3.39, 3.51]) was significantly smaller than that of Hispanic American categories ($H_{hispanic}$ = 3.57, 95% CI = [3.53, 3.62]; see Table 4).

6 Discussion

We find consistent evidence that GPT-4 portrays females as more homogeneous than males and Asian Americans as more homogeneous than White Americans in the context of jobs. However, we did not find consistent evidence for homogeneity bias against African and Hispanic Americans. There was no significant difference in entropy measures between African and White Americans whereas we found mixed evidence of bias between Hispanic and White Americans. We provide possible explanations for the divergent findings in the section that

follows:

6.1 Divergent Findings

First, unlike previous work that had granted ChatGPT the freedom to generate text within the boundaries of text format (e.g., story, character description) and length (i.e., 30 words), data collected in this work was limited to job titles. It is possible that homogeneity bias manifests in the form beyond job titles that our data fails to capture.

Second, unlike previous work that explicitly signaled race/ethnicity and gender using group labels (e.g., “Asian American man"), this work signaled group identities using names. As Elder and Hayes (2023) note in their work, names can signal a variety of traits beyond group membership. Consequently, using names may not uniformly activate stereotypical associations that group labels might be more effective at doing. In fact, we observe that the name, “Gabriel", used to represent White American males, were often (15 out of 50) associated with “Archangel", suggesting that the name was not signaling the intended group identity and instead signaling a specific Biblical figure.

Finally, unlike previous work that used cosine similarity of sentence embeddings to quantify homogeneity in the generated text, this work used entropy. These are two distinct measures that operationalize homogeneity differently and have distinct mathematical properties. The use of entropy in this study offers a unique perspective on the variability of jobs GPT-4 associates with gender and racial/ethnic groups whereas cosine similarity of sentence embeddings offers insight into the overlap of semantic content in longer forms of text. This distinction underscores the importance of methodological choices in LLM bias research and highlights the need for complimentary approaches to fully understand the nuances of LLM behavior.

6.2 Limitations & Future Work

While we acknowledge that the text completions covered in this work is limited in scope, occupation is a widely studied area of bias in the field of NLP (e.g., Bolukbasi et al., 2016; Garg et al., 2018). Furthermore, we believe that this work introduces a framework and a benchmark measure for homogeneity bias assessment in LLM-generated text. We introduce a systematic framework for homogeneity bias assessment that can be extended to other aspects of human experience including, and not limited to, cultural practices, language use, and identity expression.

Furthermore, there are reasons to question the reliability of industry category classification. As noted in the Experiment section, we had to iteratively prompt GPT-4 to correctly classify jobs into matching industry categories. It is possible that the misclassification of jobs significantly influenced entropy measurements, considering that approximately 9.5% of completions were classified into the “Other" industry category.

7 Conclusion

This work advances our understanding of homogeneity bias in Large Language Models (LLMs) by introducing entropy as a novel measure for assessing the diversity of LLM outputs in the context of job titles and industry categories. Our findings reveal more homogeneous portrayals of females and Asian Americans compared to males and White Americans, highlighting the subtle yet significant ways biases are embedded in LLM-generated content. This emphasizes the necessity for continuous scrutiny and improvement of these models to ensure fair and diverse representations.

Looking forward, the methodological approach and insights gained from this research provide a foundation for further explorations into homogeneity bias in LLMs. Future studies should aim to expand this framework to a wider array of human experiences and refine classification methodologies to enhance the accuracy of bias measurement. As LLMs increasingly influence social discourse, it is imperative to develop and maintain systems that are both equitable and representative of the diverse fabric of society.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, FAccT '21, pages 610–623. https://doi.org/10.1145/3442188.3445922.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *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.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, NIPS'16, pages 4356–4364.

Elizabeth Mitchell Elder and Matthew Hayes. 2023. Signaling Race, Ethnicity, and Gender with Names: Challenges and Recommendations. *The Journal of Politics* 85(2):764–770. Publisher: The University of Chicago Press. https://doi.org/10.1086/723820.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115(16):E3635–E3644. Publisher: Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1720347115.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, editors, *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy, pages 166–172. https://doi.org/10.18653/v1/W19-3823.

Messi Lee, Jacob Montgomery, and Calvin Lai. 2023. The effect of group status on the variability of group representations in LLM-generated text. In *Socially Responsible Language Modelling Research*. https://openreview.net/forum?id=izC60DLg29.

David J.C. MacKay. 2005. *A Brief History of Time: From the Big Bang to Black Holes*. Cambridge University Press, Cambridge.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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, pages 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, pages 3982–3992. https://doi.org/10.18653/v1/D19-1410.

A Appendix

A.1 Industry Categories

1. Healthcare Industry
2. Technology and IT Industry
3. Education and Academic Industry
4. Financial Services Industry
5. Manufacturing and Engineering Industry
6. Retail and Commerce Industry
7. Hospitality and Tourism Industry
8. Transportation and Logistics Industry
9. Construction and Real Estate Industry
10. Creative Arts and Entertainment Industry
11. Agriculture and Forestry Industry
12. Energy and Utilities Industry
13. Legal Services Industry
14. Environmental and Conservation Industry
15. Government and Public Administration
16. Media and Communications Industry
17. Non-Profits and Social Services Industry
18. Translation Services Industry
19. Sports Industry
20. Other