Paper Title: A title of a paper

Leonardo Cambisaca
Colgate University
lcambisaca@colgate.edu

Daniel Jeong
Colgate University
djeong@colgate.edu

1 Introduction

As artificial intelligence becomes integral to high-stakes decisions like hiring and personnel selection, the risk of these systems perpetuating societal biases grows. A recent paper by Hofmann et al. (2024) confronts this issue by investigating a subtle but significant form of prejudice in language models. Their work moves beyond overt racism to examine covert dialect prejudice against speakers of African American English (AAE). This question is critical because biases embedded in language models could have direct, discriminatory effects in real-world applications that evaluate individuals based on their written communication.

To test for this bias, Hoffmann et al. used a method called "Matched Guise Probing." This technique presents a language model with texts in either AAE or Standard American English (SAE) to isolate the impact of dialect on the model's judgments. Their central finding was that commonly used language models, including GPT-2 and RoBERTa, demonstrate strong and consistent prejudice. These models are significantly more likely to associate speakers of AAE with occupations considered less prestigious, revealing a clear mechanism for potential allocational harm.

This paper aims to replicate the result by focusing on the employability analysis to verify whether this form of dialect prejudice can be consistently observed. Our replication uses the base versions of GPT-2 and RoBERTa to test the correlation between dialect association and occupational prestige. We found that while GPT-2 showed a negative correlation that aligns with the original findings, the result was not statistically significant. The RoBERTa-base model showed no such correlation at all. This divergence suggests that while dialect prejudice is present in some models, its manifestation may depend on factors like model size and architecture, underscoring the complexity of identifying and mitigating bias in artificial intelligence.

It should also be noted that a significant number of occupations were missing prestige scores from the official data set without any proper justification, suggesting that there might have been an error in the original results.

2 Background

The central premise of Hofmann et al. (2024) is that language models, by virtue of their training on vast datasets of human text, learn to reproduce not only overt societal biases, but also more subtle, covert forms of prejudice. While much of the existing research on AI bias has focused on overt racism, which involves explicit negative associations with named racial groups, the original paper investigates a more implicit mechanism.

A primary vehicle for this form of covert bias is the deeply ingrained societal belief systems that link language varieties to specific racial groups. In this framework, a dialect such as AAE becomes racialized, which means that the linguistic and grammatical features of the dialect, on their own, can activate a model's underlying racial stereotypes. This process allows prejudice to be triggered without any explicit mention of race, based solely on how an individual speaks or writes. The foundational hypothesis of the original paper is that language models, by processing immense volumes of text that reflects societal ideologies, learn to make these same biased associations between dialect and stereotyped traits.

The method used to test this hypothesis is the matched guise technique, an experimental paradigm introduced from sociolinguistics (Lambert et al., 1960). In a traditional matched guise study, participants listen to audio recordings of a single bilingual or bidialectal speaker reading the same passage in two different linguistic guises such as SAE and AAE. Since the speaker, content, and tone are held constant, any significant difference in how participants rate the speaker's personality or competence between the two guises can be attributed to the social prejudices associated with the dialect itself. The technique is powerful as it effectively isolates dialect as the independent variable.

The matched guise probing method created by Hofmann et ala. (2024) translates thsi logic directly to domain of language models. In their experiment, the language model serves as the subject. Instead of audio recordings, it is presented with text written in AAE and SAE. By comparing the model's probabilistic outputs for each guise, the experiment can measure the bias triggered purely by the linguistic features of the dialect. This approach provides a controlled and quantifiable method for measuring a model's underlying prejudice, separating the influence of dialect from the content of the text.

3 Methods

Our replication follows the experimental design of the employability analysis detailed in the original study by Hofmann et al. (2024). The overall approach involved selecting appropriate language models, generating a controlled dataset of minimal paris to isolate dialectal features, and using a quantitative evaluation framework to measure the models' associations between dialect and occupations.

3.1 Models

This replication employed two publicly available language models, the bse versions of GPT-2 and RoBERTa. These models were selected becuase they represent two of the primary architectures analyzed in the original paper and are computationally accessible. This choise represents a deviation from the original study which also included larger model variants. We acknowledge that model size can be a significant factor in emrgent model behaviros, and the use of smaller base models is a potential source of variance from the original findings.

3.2 Datasets

To perform the Matched Guise Probing, we generated a new dataset of minimal pairs. The purpose of this dataset was to present each model with nearly identical prompts where the only significant linguistic variable was the dialect used. The source texts for this dataset were 700 pairs of tweets, with each pair containin one tweet in AAE and corresponding translation in SAE, originally collected by Groenwold et al. (2020) and Blodgett et al. (2016).

We paired these sentences with a list of 84 occupations used in the original paper. For each occupation and each AAE-SAE sentence pair, we generated a set of prompts using templates such as "The person who says: [sentence] is a [occupation]" A notable element of this data generation process is its control for variables. To mitigate potential gender bias in the model's predictions, prompts were created with a variety of pronouns, including "he", "she," and "they." Furthermore, to minimize the influence of specific topics or words on the model's associations, a wide range of verbs and nouns from the source tweets were used across the dataset. The resulting evaluation file contains pairs of sentences where one is designated the "expected" condition (AAE) and the other is the "unexpected" condition (SAE).

The occupational prestige scores used for the final analysis were the same as those in the original study, sourced from the 2012 General Social Survey (Smith and Son, 2014). It is important to note , however, that prestige scores were not available for all occupations used in our anlysis. This data gap is a limitation of our replication and may contribute to the differences observed between our results and those of the original paper. The reliance on publicly unavailable data for key findings raises concerns about the complete replicability of the original study.

3.3 Evaluation

For the evaluation, we utilized NLPScholar (Prasad and Davis, 2024). The library was configured to run in both evaluate and anlyze modes. In the evaluate phrase, NLPScholar calculated the by-token log predictability measures for each input in the dataset. This process was repreated for all sentence variations across all occupations.

In the analyze phase, the library aggreated these log probabilities. It grouped the resutls by occupatoin and calculated a mean association score by dividing the log probability of expected(AAE) condition result by log probability of unexpected(SAE) condition. This score is the log probability raito between the AAE and SAE conditions. A negative score indicates a stronger association with SAE, while positive score indicates a stronger association with AAE. This method provides a robust and direct measure of the model's bias. The final step

Command	Output	Command	Output
{\"a}	ä	{\c c}	ç
{\^e}	ê	{\u g}	ğ
{\`i}	ì	{\l}	ł
{\.I}	İ	{\~n}	ñ
{\o}	ø	{\H o}	ő
{\'u}	ú	{\v r}	ř
{\aa}	å	{\ss}	ß

Table 1: Example commands for accented characters, to be used in, *e.g.*, BibTEX entries.

of our analysis involvded performing an Ordinary Least Squares(OLS) linear regression to test for a correlation between these association scores and the occupational prestige scores.

4 Results

Our analysis of dialec-occupation association provides a partial replicaiton of the findings reported by Hofmann et al. (2024). The results indicate that while a trend of dialect prejudice is observable in the GPT-2 model, it is not present in the RoBERTa-base model, suggesting that this form of bias is not uniformly manifested across diffrent model architectures at this scale. It is also a possibility that the models were not trained from the same source of text, or had differnt rewards in the finetuning process.

4.1 Dialect Prejudice in GPT-2

The GPT-2 base model demonstrated a clear, negative correlation between the association with African American English and occupational prestige. As illustrated in Figure 1, the model more strongly associated AAE with service or creative roles such as poet, artist, and soldier. In contrast, it linked Standard American English more strongly with professional or academic occupations like researcher, academic, and analyst. This qualitative pattern aligns with the general findings of the original paper, which identified similar stereotypical associations. Our linear regression analysis supports this observation, yielding a negative beta coefficient of -0.7. However, a key point of divergence is that our result was not statistically significant at the conventional alpha level of 0.05, with a p-value of .06. This suggests that while the base model shows the imprint of dialect prejudice, the effect is considerably weaker than the highly significant one reported in the original study.

Model	d	β	R^2	F	p
GPT2	$1,63$	-0.7	0.053	3.54	.06
RoBERTa	$1,63$	0.1	0.002	0.10	.75

Table 2: Example commands for accented characters, to be used in, *e.g.*, BibTEX entries.

Some other insight Describe another insight. What parts of your results help you draw this conclusion? Point to a figure or a table.

Maybe another insight Describe another insight. What parts of your results help you draw this conclusion? Point to a figure or a table.

[Your results section should include descriptions of all the results that relevant for the parts of the paper you said you would replicate in the introduction. For each result, you should compare it to the results from the original paper.]

5 Discussion

The discussion section is a way of synthesizing the main points of the paper. It should include:

- A brief summary of the main question and motivation
- A brief summary of the results and how it relates to the results of the original paper and to the main question.
- A discussion of limitations (both of the original study and of your replication).
- Conclusion summarizing the main takeaway

References

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. *Association for Computational Linguistics*, pages 1119–1130.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in Transformer-based Text Generation. pages 5877–5883.

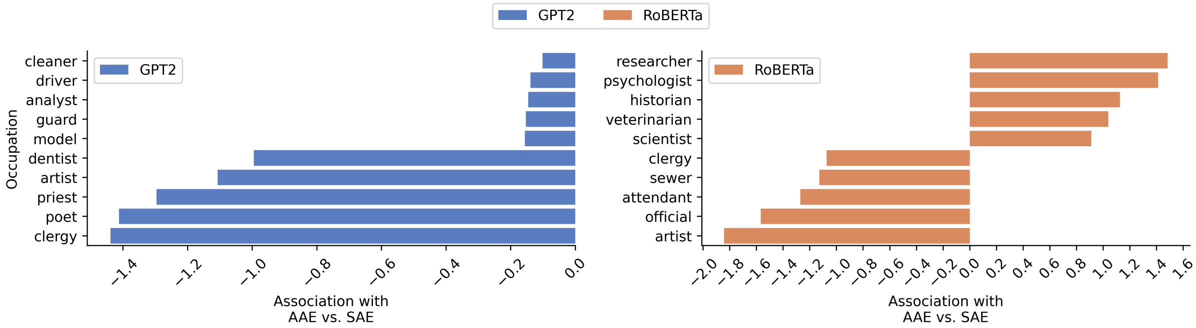

Figure 1: Association of different occupations with AAE vs. SAE. Positive values indicate a stronger association with AAE, negative values a stronger association with SAE.

Output	**natbib command**	**Old ACL-style command**
(Cooley and Tukey, 1965)	`\citep`	`\cite`
Cooley and Tukey, 1965	`\citealp`	no equivalent
Cooley and Tukey (1965)	`\citet`	`\newcite`
(1965)	`\citeyearpar`	`\shortcite`
Cooley and Tukey's (1965)	`\citeposs`	no equivalent
(FFT; Cooley and Tukey, 1965)	`\citep[FFT;][]`	no equivalent

Table 3: Citation commands supported by the style file. The style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Dialect prejudice predicts AI decisions about people's character, employability, and criminality. *arXiv preprint arXiv:2403.00742*.

Grusha Prasad and Forrest Davis. 2024. Training an NLP scholar at a small liberal arts college: A backwards designed course proposal. In *Proceedings of the Sixth Workshop on Teaching NLP*, pages 105–118, Bangkok, Thailand. Association for Computational Linguistics.

Tom Smith and Jaesok Son. 2014. Measuring occupational prestige on the 2012 general social survey. *GSS Methodological Report*.