Dialectic Preference Bias in Large Language Models

Muhammad Furquan Hassan[1], **Faiza Khan Khattak**[1, 2], **Laleh Seyyed-Kalantari**[1,3]

[1]York University, 4700 Keele St, North York, ON M3J 1P3, Canada
[2]Monark Health, Toronto, Canada
[3]Vector Institute, 108 College St W1140, Toronto, ON M5G 0C6, Canada
hfurquan, lsk@yorku.ca, faizakk@monark-health.com

Abstract

Dialectic preference is an often overlooked language model's (LLM) bias against marginalized groups. It can be observed When LLMs output reflects or promotes unfair preferences or prejudices towards particular dialects or linguistic variations. Such bias may lead the model to favor certain ways of speaking or writing, which can disadvantage speakers of marginalized dialects. Such bias can perpetuate social biases and inequalities, affecting how people interact with and are supported by AI technologies. In this preliminary study, we analyze dialectic preference bias for Standard American English (SAE) compared to African American English (AAE) using the sentiment classification task on Claude 3 Haiku, Phi-3-medium, and LLaMa 3.1 8b. We find a greater tendency to classify AAE sentiments as negative, especially in LLaMa 3.1 8b, compared to other models, demonstrating the presence of dialectic preference bias. This work highlights the importance of addressing dialectic language-based biases in LLMs to reach inclusive and equitable LLMs. We plan to extend this study to more dialects and larger language models.

Introduction

The intersection of sociology, linguistics, and artificial intelligence (AI) has never been more apparent than in today's rapidly evolving landscape of language models (LLMs) (Brown et al. 2020). These sophisticated AI models, seamlessly integrated into our daily lives through chatbots and automated customer service, are reshaping how we interact with technology and, by extension, with language itself. However, as these models become more pervasive, they mirror and amplify the societal biases embedded in their training datasets (Gallegos et al. 2024).

One often overlooked bias is *dialectic preference bias*. This phenomenon occurs when language models (LLMs) exhibit systematic asymmetries in their outputs, reflecting and potentially reinforcing societal prejudices in preference of or against certain dialects or linguistic variations. Such biases are particularly insidious because they operate at the level of word choice and phrasing, subtly shaping perceptions and reinforcing social-category cognitions about the described individuals or groups.

This preliminary study focuses on the dialectic preference bias between Standard American English (SAE) and African American English (AAE). We aim to uncover and quantify dialectic preference bias of three LLMs—*Claude 3 Haiku*, *LLaMa 3.1 8b*, and *Phi-3-Medium*—on sentiment classification tasks, while using *GPT-4o-mini* as a translator of AAE dialectic to SAE and SAE to AAE. Our method involves creating parallel datasets in SAE and AAE, allowing for a direct comparison of model performance across these linguistic variations. Our contributions are as follows:

- We introduce, define and demonstrate the concept of **Dialectic Preference Bias** for LLMs.
- We introduce the **Dialectic Group Invariance** (DGI) metric to quantify the LLMs'invariance to dialectal variations.
- We release a collection of parallel SAE-AAE datasets **datasets**[1], offering the research community a valuable resource for investigating and mitigating linguistic biases.

Related Work

The study of bias in LLMs has gained significant attention since the success of GPT-3 (Brown et al. 2020). Researchers have explored various types of biases, including gender, racial, and religious biases, and their impacts on model performance and fairness (Gallegos et al. 2024; Kohankhaki et al. 2024; Tian et al. 2023). Su Lin Blodgett (Blodgett 2021) examined language identification and modelling for African American English (AAE) on Twitter, highlighting challenges due to linguistic variation and the misclassification risk. They emphasize the necessity of evaluating and understanding bias against AAE. The datasets of AAE tweets released under this study serve as the basis for our study. Patel and Pavlick (Patel and Pavlick 2021) explored the subtleties of linguistic framing effects on generative language models and found that models are sensitive to bias-inducing linguistic markers like hedges and assertives. Their showed even minor stylistic elements in prompts can influence the generated content's polarity and subject matter. The concept of linguistic bias, as defined by linguistic anthropologists Beukeboom and Burgers (Beukeboom and Burgers 2019), serves as a theoretical underpinning for our study. Their work on how stereotypes are

[1]https://huggingface.co/datasets/furquan/dialectic-preferences-bias-aae-sae-parallel

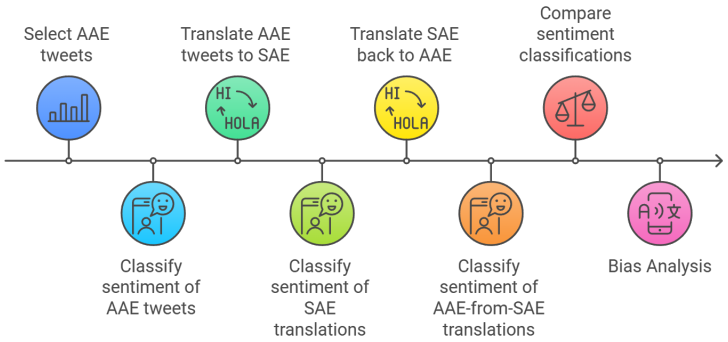


Figure 1: The complete pipeline for this study.

shared through language, particularly the notion of systematic asymmetries in word choice reflecting social-category cognitions, directly relates to our investigation of *dialectic preference bias*.

Methodology

Our study employs a multi-faceted approach to investigate dialectic preference bias in three prominent LLMs: Claude 3 Haiku, Phi-3-Medium, and LLaMa 3.1 8b. We utilize sentiment analysis and dialect translation (using *GPT-4o-mini*) to quantify potential biases of different LLMs in these tasks.

Dataset Preparation

We began with the AAE tweet dataset released by Su Lin Blodgett (Blodgett 2021), selecting 5000 tweets containing more than 10 words to ensure meaningful content. The dataset was cleaned to remove non-ASCII characters, providing a robust foundation for our analysis.

Sentiment Classification and Dialect Translation

We conduct experiments using two closed-source language models, GPT-4o-mini and Claude 3 Haiku, as well as two open-source models, LLaMA 3.1 (8B) and Phi-3-Medium. Given the sentiment function $f : X \rightarrow \{-1, 0, 1\}$ (-1 is negative, 0 is neutral and 1 is positive) for each language model, we performed the following steps:

1. Obtain sentiment classifications for the original AAE tweets i.e., $f(x^i_{AAE}) \ \ \forall i$ tweets.
2. Translate AAE tweets to Standard American English (SAE) using the *GPT-4o-mini*.
3. Obtain sentiment classifications for the SAE translations i.e., $f(x^i_{SAE}) \ \ \forall i$ tweets.
4. Translate the SAE sentences back to AAE (using *GPT-4o-mini*) to assess the model's perception of AAE.
5. Obtain sentiment classifications for the AAE-from-SAE translations i.e., $f(x^i_{T(AAE)}) \ \ \forall i$ tweets.

Ideally, $f(x^i_{AAE})$, $f(x^i_{SAE})$, and $f(x^i_{T(AAE)})$ should be identical, indicating the absence of dialectical bias.

Dialectic Group Invariance (DGI) Metric

We define the Dialectic Group Invariance (DGI) metric to quantify the models' bias across different dialects. For N sentences, DGI is defined as:

$$DGI(X_{AAE}, X_{SAE}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(f(x^i_{AAE}) = f(x^i_{SAE})) \tag{1}$$

here, the $\mathbb{1}$ indicator function equals 1 if the condition holds, otherwise 0. DGI $\in [0, 1]$, with a higher value indicating less bias and sentiment shift (i.e., consistent behavior) across dialects, AAE and SAE. We also calculate the two-way DGI across AAE, SAE, and Tr(AAE) using Eq.(2).

$$DGI(X_{AAE}, X_{SAE}, X_{T(AAE)}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(f(x^i_{AAE}) = f(x^i_{SAE}) = f(x^i_{T(AAE)})) \tag{2}$$

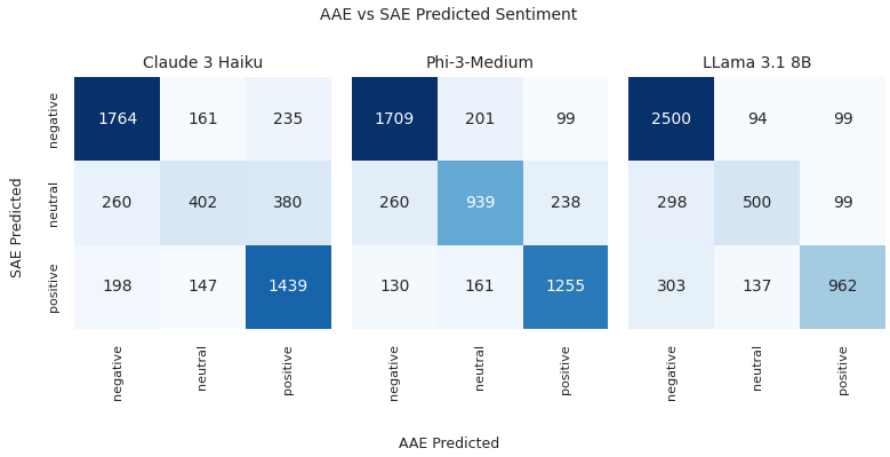
Result and Discussion

DGI Score Analysis

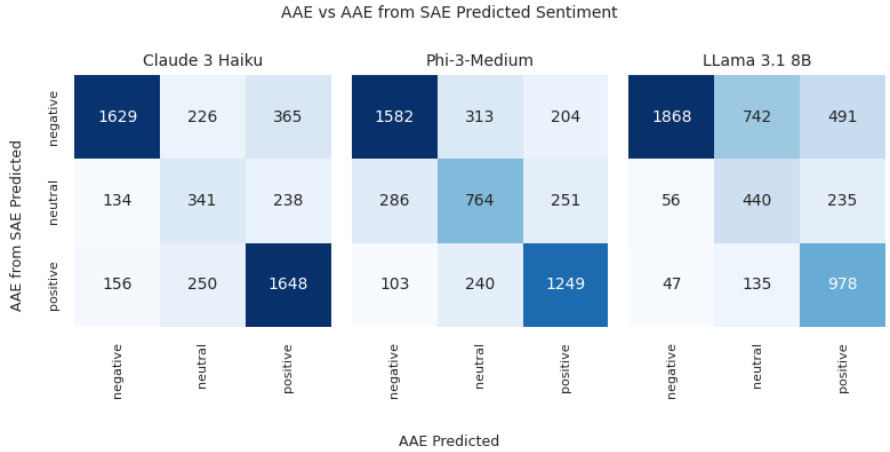
The DGI scores reported in 1 show inconsistent sentiment classifications of the language models for semantically equivalent AAE and SAE inputs for 21-28%. The inconsistency is even higher at 29-44% if we consider a two-way analysis showing the translation and re-translation sentiment shift. Such an outcome highlights dialectic preference bias, showing that even state-of-the-art models struggle to maintain consistent performance across dialects.

To analyze sentiment across different dialects, we use confusion matrices with model-assigned labels for both AAE and translated SAE texts. This comparison relies on the semantic invariance of the texts, as different dialects convey the same meaning. We use the same language model for all translations to keep translation bias constant across experiments. We used GPT-4o-mini for translations because of it predominant use and easy access.

2a shows the confusion matrix of predicted sentiments classification results for AAE and the corresponding predicted sentiment classification of SAE translation of the same sentences. Here, are our observations:



(a) Sentiment assigned to AAE vs same sentences translated to SAE.



(b) Sentiment assigned to AAE vs AAE obtained by translating SAE.

Figure 2: Confusion Matrices for Clause 3 Haiku, Phi-3-Medium and Llama-3.1-8B with GPT-4o-mini as the translation model

Model	DGI (X_{AAE}, X_{SAE})	DGI (X_{AAE}, X_{SAE},$X_{T(AAE)}$)
LLama 3.1 8B	0.79	0.67
GPT-4o-mini	0.79	0.71
Phi-3-Medium	0.78	0.66
Claude-3-Haiku	0.72	0.67

Table 1: All models failed to achieve a DGI score of 1.0. The lower the DGI score the more dialectic bias. The First column indicates the DGI score between AAE and SAE, while the second column indicates the two-way DGI score of AAE vs AAE obtained from translating SAE back to AAE.

A) Bias in negative sentiment assignment for AAE: In the confusion matrix 2a, among the three models and the same set of sentiments, LLama 3.1 assigns more negative 3,101 instances (i.e., sum of the all instances in the negative column of AAE 2500+298+303) to AAE, where Phi-3-Medium and Claude 3 Haiku assign 2,099 and 2,222 instances, respectively. Note that the test set is the same and only models are different. Additionally, among the three models, Claude 3 Haiku assigns more positive to AAE (2,054) vs LLama 3.1 (1,160) positive assignment. Llama 3.1 8B shows a significant over-assignment of negative sentiment for AAE compared to the other models, which could reflect a negative bias toward AAE sentiments, however it also turns the most negative AAE sentiment into positive SAE sentiment.

B) Bias in positive sentiment assignment for AAE: In the confusion matrix 2a, among the 3 models, Claude 3 Haiku has the most positive cases for AAE (35 + 380 + 1439 = 2054) while Phi-3-Medium and Llama 3.1 8B are in the second and third place with total of 1592 and 1160 positive assignment to AAE. This suggests that Claude 3 Haiku is more likely to assign AAE cases as positive overall than the other models, demonstrating a potential bias toward positive sentiment assignment for AAE.

C) Bias in positive sentiment assignment for SAE of an original negative AAE: LLama 3.1 gives positive to the SAE version of the originally negative in AAE (303 instances). These numbers are 198 and 130 in Claude 3 Haiku and Phi-

AAE	SAE
Relationships r wut u make it so dont be such a bitch about thing and relationships will last longer.	Relationships are what you make of them, so don't be so difficult about things and relationships will last longer.
I like him too much to stop buht I LOVE him way too much to keep goin I think I kno wat I gotta do	I like him too much to stop, but I love him way too much to keep going. I think I know what I have to do.
Not goin to sleep til she wake dat ass up and call me so until then..... #teamupalldamnnight!!!	I am not going to sleep until she wakes up and calls me, so until then... #teamupalldamnnight!!!
This man supposed to buy my car tomorrow, i hope he do i HATE driving a stick.	This man is supposed to buy my car tomorrow. I hope he does; I hate driving a manual transmission.

Table 2: Examples of translation of AAE sentences to SAE using GPT-4o-mini

Label	Sentence	Sentiment
AAE	Thank you Lord for waking me up this morning ...please give me the strength cuz these ppl irking my nerves.	Negative
SAE	Thank you, Lord, for waking me up this morning. Please give me the strength because these people are getting on my nerves.	Positive

Table 3: Example Sentence Misclassified by Claude 3 Haiku

3-Medium, respectively. This shows the bias of LLama 3.1 in analyzing SAE as positive.

D) Bias in negative sentiment assignment to translated AAE: The analysis of 2b, comparing sentiment analysis of AAE and T(AAE), demonstrate that Llama 3.1 has a higher tendency to assign negative sentiment to T(AAE) examples (491 cases) that were originally positive in AAE, showing more bias toward negative assignment in handling positive sentiment during the translation process or sentiment analysis. This number is the least (204) for Phi-3-Medium.

E) Re-translation sentiment shift: 2b shows that Llama 3.1 assigns 742 neutral AAE sentiments to negative after translation. However, only 94 neutral AAE instances were assigned negative in SAE (from 2a). It shows a drastic sentiment shift to negative. Llama 3.1 is especially prone to over-assigning negative sentiment during the re-translation of SAE to AAE. Surprisingly, negative sentiments are less likely to shift to neutral or positive compared to positive shifting to negative or neutral.

F) Positive AAE sentiments remaining positive after translation: 2b shows Llama 3.1 has the lowest number of positive AAE sentiments remaining positive (978) while Claude 3 Haiku retains the highest number (1648).

Qualitative analysis: We also performed a qualitative analysis of misclassified examples, revealing that cultural context and dialect-specific expressions often cause sentiment misinterpretation. For example, 3 sentence with the phrase "please give me the strength cuz these ppl irking my nerves" was classified as negative in AAE but positive in SAE, highlighting the challenge of accurately interpreting culturally-specific language.

Discussion and Conclusion

We show strong dialectic bias against AAE across three models, with Llama 3.1 having the most substantial bias and Claude 3 Haiku having the least. We observed bias amplification during the re-translation process, with the highlights systematic sentiment shifts during re-translation, with Llama 3.1 resulting in disproportionately shifting neutral and positive sentiments to negative. Conversely, negative sentiments are less likely to shift to neutral or positive than positive sentiments shifting to negative or neutral.

Negative sentiment amplification for AAE could lead to biased outcomes in real-world sentiment analysis applications, such as social media monitoring or customer feedback analysis, disproportionately affecting speakers of AAE.

Implications, Limitations, and Future Work

These results collectively demonstrate a dialectic preference bias in current state-of-the-art LLMs. The inconsistency in sentiment classification across dialects could lead to real-world consequences, particularly in applications involving sentiment analysis, content moderation, or any NLP task where emotional valence is important.

To the best of our knowledge, this is one of the first efforts to explore dialectical bias in language models. However, the study is limited to a few models, a small set of randomly selected 5000 samples, and a restricted set of dialects. A more thorough investigation is needed to achieve greater clarity. Future work will focus on using a diverse set of LLMs with different sizes to determine if they are also prone to dialectical bias, involving human-in-the-loop and domain experts to conduct qualitative analysis and find the root causes, expanding the analysis to a larger AAE-SSE dataset and releasing the dataset, extending the exploration to include various other dialects including the low-resource ones, developing training techniques and data curation methods that promote more equitable representations of diverse dialects and developing bias mitigation strategies that can be applied post-training.

Acknowledgments

We sincerely thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for the Discovery Grant (to Dr. Laleh Seyyed-Kalantari) and the Connected Minds CFREF (to Dr. Seyyed-Kalantari).

References

Beukeboom, C. J.; and Burgers, C. 2019. How stereotypes are shared through language: a review and introduction of the aocial categories and stereotypes communication (SCSC) framework. *Review of Communication Research*, 7: 1–37.

Blodgett, S. L. 2021. *Sociolinguistically Driven Approaches for Just Natural Language Processing*. Ph.D. thesis, University of Massachusetts Amherst.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 1–79.

Kohankhaki, F.; Tian, J. J.; Emerson, D.; Seyyed-Kalantari, L.; and Khattak, F. K. 2024. Reevaluating Bias Detection in Language Models: The Role of Implicit Norms. In *TrustNLP Workshop at the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*.

Patel, R.; and Pavlick, E. 2021. "was it "stated" or was it "claimed"?: How linguistic bias affects generative language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10080–10095.

Tian, J.-J.; Emerson, D.; Pandya, D.; Seyyed-Kalantari, L.; and Khattak, F. K. 2023. Efficient Evaluation of Bias in Large Language Models through Prompt Tuning. In *Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2023*.