Black LLMirror: User (Self) Perceptions in Black American English Interactions with LLMs

Anonymous authors
Paper under double-blind review

1 Introduction & Motivation

Popular large language models are becoming increasingly capable of generating more personalized interactions finely tuned to the needs, wants, and preferences of individual users (Zhang et al., 2025). Among these tasks involved in personalization is style imitation (Chen & Moscholios, 2024), or the act of adapting to a user's style of writing. Risks can occur, however, when an LLM mistakes a minority user's dialect for their style of writing and naively attempts to adapt accordingly. This sort of personalization may give way to harms against the user. In some such cases, an LLM's use of a minority dialect risks misunderstanding the user and subsequently bringing harm to their self-esteem (Wenzel et al., 2023; Cunningham et al., 2024), or engaging in what can be understood as appropriative behavior (Basoah et al., 2025a). Depending on the task for which the LLM is used, users may determine that the use of dialect may be unhelpful, inappropriate, or offensive (Sandoval et al., 2025). In other such cases, users may be frustrated to find that, should their desired task call for use of their dialect, LLMs are ill-equipped to generate acceptably natural language in their dialect (Basoah et al., 2025b), or to understand input given in their dialect (Koenecke et al., 2020; Mengesha et al., 2021). The potential harms of misalignment in dialect use are especially prevalent in speech-based interactions as compared to text-based interactions (Wenzel et al., 2023; Hurst et al., 2024).

In this study, we investigate user responses to a dialect-specific language model dependent on the model's modality and domain as well as use of dialect, following the findings of (Basoah et al., 2025a), which explores perceptions of an LLM's use of two closely related sociolects (Finegan & Rickford, 2004), African American Vernacular English and Queer slang. We look specifically into the responses of speakers of African American Vernacular English, referred to here as Black American English or BAE (Hall et al., 2021), and more widely into how perceptions vary across speech- and text-based interactions and across different domains of conversation. This ongoing work designs a set of user studies to determine BAE-speaking user perceptions both of a BAE-specific language model and of the users themselves after interacting with the model to be run at scale with BAE-speaking participants. We measure perceptions of the model across axes of trust (Cohn et al., 2024), understandability, and social presence (Zhou et al., 2024), and user self-perceptions across axes of general feeling (Watson et al., 1988), public self-consciousness (Fenigstein et al., 1975), and collective self-esteem (Luhtanen & Crocker, 1992).

Our research questions are as follows:

- What effects do different modalities and domains of interaction have on native BAE speakers' perceptions of a language model that produces BAE?
- What effects do different modalities and domains of interaction have on native BAE speakers' perceptions of themselves following their interaction?

We expect that, in line with the findings of Basoah et al. (2025a), BAE speakers will rate a language model that responds using Standard American English or SAE more positively, and will rate specifically higher on the axis of trust across both spoken and written interactions. Howeve, we also anticipate a considerable difference in responses between text- and speech-based interactions, and that overall, participants will enjoy their experience with the BAE-specific model more than their overall experiences with the SAE model.

2 Proposed Methodology

User Study Design Each participant will be asked to interact with a language model determined by modality (i.e. speech or text), dialect (i.e. BAE or SAE), and domain (i.e. education, small tasks, personal conversation, or impersonal conversation), so that participants will take part in one of sixteen different types of interaction as described above for the purposes of our research. Participants will be provided with prompts unique to each domain to begin interacting with the language model, but will not be restricted to the use of any of the prompts. Four prompts are provided in total for each domain. Participants will interact either with a language model that is prompted to respond to act as a helpful assistant that only responds using BAE, or with one prompted to act as a helpful assistant that uses SAE. Participants will not be made aware of the model they have been randomly assigned. Regardless, participants in both groups will be asked to interact with the model using BAE as often as possible.

Perception Variables Before and after their interactions with the language model, participants will be asked to answer an identical set of survey questions concerning themselves across axes of general feelings (Watson et al., 1988), public self-consciousness (Fenigstein et al., 1975), and collective self-esteem (Luhtanen & Crocker, 1992), the answers to which will be compared to determine whether any changes can be attributed to the interaction. All such questions are designed to be answered on a 7-point Likert scale (e.g. ”To what extent do you agree with the following statement: *I'm usually concerned about what other people think of me.*”) to easily process positive and negative sentiment for quantitative analysis. In addition, participants will also be asked to answer questions concerning their perceptions of and experiences with the language model across variables of trust (Cohn et al., 2024), understandability, and social presence Zhou et al. (2024). Likewise, these questions are also designed to be answered on a 7-point Likert scale.

BAE LLM setup We used Qualtrics to create our survey, which allows us to edit the Javascript of each survey question and include a plugin for an API key. We used ChatGPT-4o to create both text- and speech-based interactions as survey questions in their larger respective surveys. To prompt the BAE-specific models to produce BAE, we used a combination of rule-based prompting (Ziems et al., 2022) and persona prompting, as we found rule-based prompting to be insufficient to the task of generating satisfactory utterances of BAE (Sun et al., 2024). We created pre-screener surveys over both text and speech each with 8 phrases of BAE generated by ChatGPT as prompted above and 2 phrases of SAE. Participants will be asked to identify which dialect of a selection of minority dialect the phrases seem to align with. We expect that this will validate our prompting process.

Participant Recruitment Our recruitment process closely follows that of (Basoah et al., 2025a). We will recruit participants through Prolific, advertising in a short description of our study that as participants, it will be crucial to be “speakers or common users of AAVE (African American Vernacular English), or come from communities that speak AAVE.” Prolific allows us to filter for users based in the United States, who self-identify as Black American, and who are over the age of 18, which will create a survey pool of self-identified Black American adults based in the U.S. Participants are informed through Prolific that our study will take an estimated 20 minutes to complete and that they will be compensated for their time at an hourly rate of $15 for completing the survey.

3 Expected Results & Importance

Within interactions with the BAE-specific model, we expect to observe a notable difference in participant responses between text- and speech-based interactions. This would lend weight to the need for focus on speech-based minority dialect interactions with LLMs in future work. Importantly, such findings would exemplify the weight of user perceptions in performance metrics with regard to minority dialect use in LLMs.

References

Jeffrey Basoah, Daniel Chechelnitsky, Tao Long, Katharina Reinecke, Chrysoula Zerva, Kaitlyn Zhou, Mark Díaz, and Maarten Sap. Not like us, hunty: Measuring perceptions and behavioral effects of minoritized anthropomorphic cues in llms. *arXiv preprint arXiv:2505.05660*, 2025a.

Jeffrey Basoah, Jay L Cunningham, Erica Adams, Alisha Bose, Aditi Jain, Kaustubh Yadav, Zhengyang Yang, Katharina Reinecke, and Daniela Rosner. Should ai mimic people? understanding ai-supported writing technology among black users. *arXiv preprint arXiv:2505.00821*, 2025b.

Ziyang Chen and Stylios Moscholios. Using prompts to guide large language models in imitating a real person's language style. *arXiv preprint arXiv:2410.03848*, 2024.

Michelle Cohn, Mahima Pushkarna, Gbolahan O Olanubi, Joseph M Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth. Believing anthropomorphism: examining the role of anthropomorphic cues on trust in large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2024.

Jay Cunningham, Su Lin Blodgett, Michael Madaio, Hal Daumé Iii, Christina Harrington, and Hanna Wallach. Understanding the impacts of language technologies' performance disparities on african american language speakers. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 12826–12833, 2024.

Allan Fenigstein, Michael F Scheier, and Arnold H Buss. Public and private self-consciousness: Assessment and theory. *Journal of consulting and clinical psychology*, 43(4): 522, 1975.

Edward Finegan and John R Rickford. *Language in the USA: Themes for the twenty-first century*. Cambridge University Press, 2004.

Erika V Hall, Sarah SM Townsend, and James T Carter. What's in a name? the hidden historical ideologies embedded in the black and african american racial labels. *Psychological science*, 32(11):1720–1730, 2021.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14): 7684–7689, 2020.

Riia Luhtanen and Jennifer Crocker. A collective self-esteem scale: Self-evaluation of one's social identity. *Personality and social psychology bulletin*, 18(3):302–318, 1992.

Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. "i don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on african americans. *Frontiers in Artificial Intelligence*, 4:725911, 2021.

Sandra C Sandoval, Christabel Acquaye, Kwesi Cobbina, Mohammad Nayeem Teli, and Hal Daumé III. My llm might mimic aae–but when should it? *arXiv preprint arXiv:2502.04564*, 2025.

Wangtao Sun, Chenxiang Zhang, XueYou Zhang, Xuanqing Yu, Ziyang Huang, Pei Chen, Haotian Xu, Shizhu He, Jun Zhao, and Kang Liu. Beyond instruction following: Evaluating inferential rule following of large language models. *arXiv preprint arXiv:2407.08440*, 2024.

David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

Kimi Wenzel, Nitya Devireddy, Cam Davison, and Geoff Kaufman. Can voice assistants be microaggressors? cross-race psychological responses to failures of automatic speech recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2023.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen Ahmed, and Yu Wang. Personalization of Large Language Models: A Survey, May 2025. URL http://arxiv.org/abs/2411.00027. arXiv:2411.00027 [cs].

Kaitlyn Zhou, Jena D Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. Rel-ai: An interaction-centered approach to measuring human-lm reliance. *arXiv preprint arXiv:2407.07950*, 2024.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. Value: Understanding dialect disparity in nlu. *arXiv preprint arXiv:2204.03031*, 2022.