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

MIKAYLA CAMPBELL, Georgetown University, USA

JOEL MIRE, Language Technologies Institute at Carnegie Mellon University, USA

MARK DÍAZ, Research Center for Responsible AI and Human-Centered Technology at Google, USA

MAARTEN SAP, Language Technologies Institute at Carnegie Mellon University, USA

LLMs becoming increasingly personalized to users' language style raises both excitement and concerns for minority users such as Black American English (BAE) speakers. Yet, previous work has predominantly focused on user perceptions of out-of-context BAE statements by LLMs rather than naturalistic multi-turn interactions, and has ignored such systems' effects on users' self-perception. In this work, we examine the effects that multi-turn interactions with speech and text BAE-producing LLMs have on BAE speakers' perceptions of the LLM and of themselves. We observe a significant change in participant self-esteem following the interactions, and notable qualitative differences between BAE-LLM and Standard American English (SAE) LLM interactions. We also observe significant effects of BAE-usage on user perception of the model within speech-based interactions. Our findings suggest that the effects of BAE-usage by an LLM agent on model- and self-perception among BAE-speaking users are complex and widely varied.

CCS Concepts: • **Natural language processing**; • **Cultural characteristics** ; • **Human-computer interaction**; • **Race and ethnicity**;

Additional Key Words and Phrases: Natural Language Processing, Large Language Models, African American Vernacular English, User Experience, User Perception, Dialect

ACM Reference Format:

Mikayla Campbell, Joel Mire, Mark Díaz, and Maarten Sap. 2026. Black LLMirror: User (Self) Perceptions in Black American English Interactions with LLMs. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3772318.3791111

1 INTRODUCTION

Popular large language models (LLMs) are becoming increasingly capable of generating more personalized interactions finely tuned to the needs, wants, and preferences of individual users [71]. Among these personalization tasks is style imitation [9], which is adapting to a user's style of writing. Problems arise, however, when an LLM mistakes a linguistic minority user's dialect for mere style and naively attempts to adapt accordingly, as though the former were the latter. In the case of Black American English (BAE) [28],[1] this sort of personalization may give way to one of two

[1] The term *African American Vernacular English (AAVE)* is used in participant-facing elements of the study in place of the term *Black American English (BAE)* as is used throughout this paper to avoid any possible interpretation of disrespect or lack of regard for the very demographic we seek to recruit for our study. While opinions regarding the labels *African American* and *Black American* vary for each individual of the ethnic demographic the labels

Authors' addresses: Mikayla Campbell, mrc195@georgetown.edu, Georgetown University, Washington, D.C., USA; Joel Mire, jmire@andrew.cmu.edu, Language Technologies Institute at Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; Mark Díaz, markdiaz@google.com, Research Center for Responsible AI and Human-Centered Technology at Google, New York, New York, USA; Maarten Sap, msap2@andrew.cmu.edu, Language Technologies Institute at Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

sorts of outcomes for the user. Depending on the task for which the LLM is used, users may deem the use of BAE unhelpful, inappropriate, or offensive [57]; may interpret its use as appropriative behavior [1]; or on account of being misunderstood by such a model [13, 41, 49], may suffer lowered self-esteem [67]. On the other hand, BAE-speaking users may be frustrated to find that LLMs are ill-equipped to generate acceptably natural language in BAE [2]. The potential harms of under-performance in dialect-specific tasks are especially prevalent in speech-based interactions as compared to text-based interactions [34, 67].

In this work, we examine this tension between personalization, usability, and appropriation with BAE as a case study. The risks involved in approaching these tasks are especially heightened in the case of BAE, due to its deep historical and cultural significance both within and beyond Black American communities, discussed at length in §2.1. Responses of BAE speakers towards language technologies have been widely varied, with some lamenting their inability to take on tasks using or involving BAE, and others expressing discomfort with language technologies that attempt to use BAE only to do so awkwardly or offensively. Prior research has investigated language technologies for their shortcomings with BAE over various tasks, including automatic speech recognition [8, 41, 43], speaker characterization [4, 30], reward modeling [51], and toxic language detection [58], the potential consequences of which range from online censorship of BAE to prejudice and discrimination in housing, employment, and sentencing against BAE speakers. Other works raise the question of whether language technologies should be designed for BAE-specific tasks at all [2, 57], finding that the risks weigh heavily on the potential benefits. Previous work [1, 21, 57], however, has focused most often on BAE speakers' judgments of language technologies in relation to BAE, as opposed to user experience as measured by participant self-perceptions given the language technology in question. In addition, these works have done so using static or single-turn statements produced by an LLM, as opposed to generating a multi-turn interaction.

In this study, we investigate BAE speakers' reactions following interactions with BAE-producing LLMs in either text-based or voice-based conversations, focusing on both participant attitudes towards the LLM (model perceptions) and towards themselves (self-perceptions). We specifically ask the following research questions:

- RQ1: What effects do different modalities of interaction (voice vs. text) have on native BAE speakers' perceptions of themselves following their interaction with a BAE-speaking model?
- RQ2: What effects do different modalities of interaction have on native BAE speakers' perceptions of a language model that produces BAE?
- RQ3: Do different dialects have an effect on participants' perceived need to code-switch when interacting with the model?

To explore these, we design two LLM interaction studies: a voice-based study where participants interact with either an LLM prompted to produce realistic-sounding, spoken BAE or an LLM that produces standard American English (SAE) speech, and a text-based study where participants type their interactions with a BAE- or SAE-producing text-based LLM. We measure perceptions of the model across axes of trust [11], understandability, and social presence [72], and user self-perceptions across axes of general feeling [66], public self-consciousness [20], and collective self-esteem [46].

Our results show slight differences in perceptions within modality between BAE and SAE interactions, displaying a negative effect of dialect usage on model perceptions within speech-based interactions. We also observe a significant effect of model dialect use on participant dialect use in both text- and speech-based interactions. Furthermore, we find that participants' open-ended commentary on the models is analytically enriching, due to its wide variation from

describe, the former has been found to invoke more neutral or positive themes than the latter [28]. To ensure that participants are treated with the utmost respect throughout the study, we chose the former because it is more widely accepted as professional. However, in an attempt to elevate and destigmatize the latter of the two titles, we refer to the dialect here as Black American English (BAE).

passionately positive to remarkably negative sentiment, encouraging integration of open-text response opportunities for participants in future similar user-experience studies. Our findings highlight the wide range of diverse reactions toward BAE-producing language technologies and the role of modality therein, contributing to a more informed approach in the design and use of language technologies for linguistic minority populations.

2 BACKGROUND & RELATED WORK

To contextualize the relevance of our study, we provide a brief overview of the BAE dialect from a sociological perspective, including historical background and its modern cultural relevance. We also address the state of BAE use and functionality by and within various language technologies, and BAE speakers' experiences with them.

2.1 Background: Black American English

Black American English, also often referred to as African American Vernacular English (AAVE) [69] or African American Language (AAL) [19], is a dialect believed to have emerged in the late 1600s, concurrent with Caribbean and other North American creoles, primarily in the Southeastern United States [52], most specifically in the Chesapeake Bay and coastal Carolina colonies [69]. Due to its region of origin, BAE is often conflated with Southern White American Vernacular English, or SWAVE. BAE itself is also expressed in a number of regional variations [24, 31], as its original speakers and their immediate descendants moved from rural Southeastern U.S. states, where enslaved Black Americans lived for centuries in considerable concentration, to Northern and Western states during World War I, following an increased need for labor, in what became known as the Great Migration, which continued through the 1960s [19, 52]. Historically, BAE has mistakenly been regarded as little more than a collection of misused or incorrect SAE [54], but the dialect is rule-governed and is expressed using its own unique and mutually intelligible phonological, morphosyntactic, and semantic phenomena [24], which supports its validity as a dialect all its own.

Historically, there has been considerable stigma surrounding the use of BAE, both by native BAE speakers and non-members of Black American communities. Native BAE speakers have faced disproportionate difficulty securing housing [47], risked being misunderstood in criminal justice proceedings [55], and have been on the bitter end of wage inequality as compared to SAE-speaking coworkers [26] on account of their dialect use. BAE can also be mistaken as obscene by non-BAE speakers [62]. Such widespread discrimination and negative perception, widely known among native BAE speakers, has in many cases discouraged the use of BAE in certain contexts, and in turn sharpened the practice of *code-switching* among speakers. Code-switching, the changing of one's language or dialect, has been described from a psychological perspective as a linguistic tool in the larger task of managing one's outward appearance or behavior [48], and efforts on the part of native BAE speakers to tailor their speech and thus avoid racial discrimination has been linked to added stress in academic and professional settings [39]. In light of these stressors, the use of BAE among non-members of Black American communities for gain in popular media [7] and politics [42] has also been widely stigmatized by native BAE speakers.

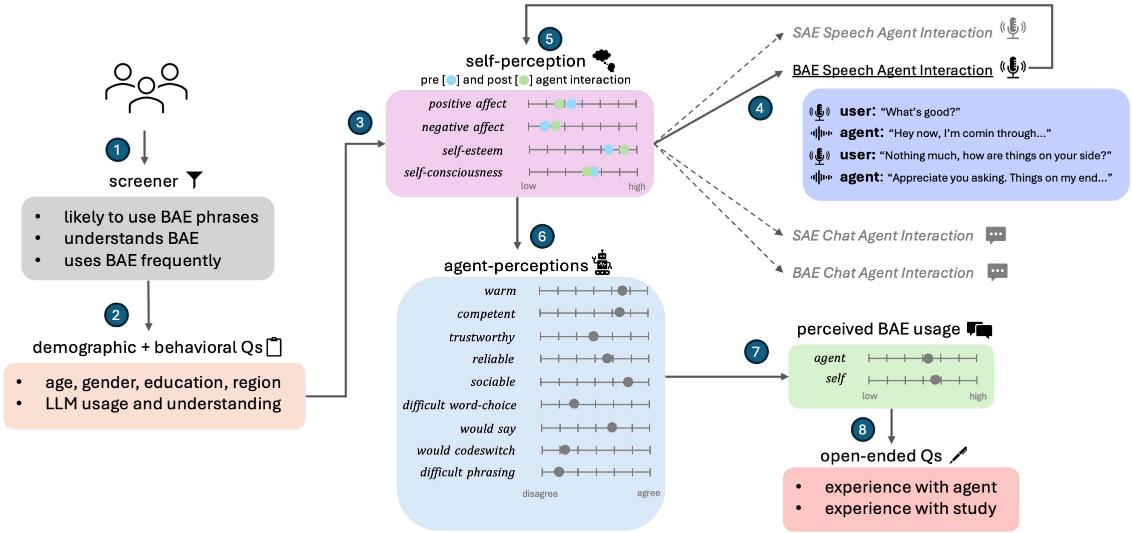
2.2 LLMs, BAE, and anti-BAE Biases

Biases against dialects in LLMs and NLP systems. A growing body of work has shown that NLP systems and large language models systematically underperform on non-standard dialects and regional varieties of English relative to Standard American English (SAE), reflecting biases embedded in training data, benchmarks, and evaluation practices [18, 27]. Such performance gaps have been documented across tasks including parsing, sentiment analysis, reasoning, translation, and speech recognition, and across English varieties such as Indian English, Caribbean English, Irish English, and other World Englishes [18, 22]. Beyond accuracy disparities, recent studies show that LLMs may misinterpret pragmatic intent, generate patronizing or corrective responses, or reinforce stigmatizing stereotypes when interacting with dialectal inputs [22]. Together, this literature suggests that dialectal bias in language technologies is a structural and widespread phenomenon [27]. Yet, every dialect—especially if it is also a sociolect—has a specific relationship to the dominant language variety, which makes the effects and harms of dialect-based biases dependent on the dialect and speaker group.

Anti-BAE biases in LLMs and NLP systems. Various previous works have documented anti-BAE biases in LLMs and NLP technology. Some works highlight key differences between BAE as spoken between native speakers and BAE found in the many corpora on which many popular LLMs are trained [3, 17], finding that the publicly accessible written BAE in popular media and on social media is insufficient to enable LLMs to acceptably imitate BAE. Instead, it often causes them to generate awkward or inappropriate utterances in BAE. An adjacent critique is posed in Finch et al. [21], which found that, in employing popular LLMs for the task of translating SAE text to both text and speech in BAE, phonetic features of BAE were generated at a disproportionate rate compared to morphological, syntactic, and semantic features, to which researchers attributed low performance across human evaluation metrics. Other works examine how well NLP approaches perform with such tasks as text generation, [15], sentiment analysis [25], and reasoning [72], the last study of which found that questions posed to a set of models in SAE were consistently answered by the models with more accurate outputs than questions posed to the same models in BAE. These findings showcase that a user's employment of BAE with many popular language technologies often yields an unsatisfactory experience as compared to the employment of SAE.

User Experiences of BAE-producing AI systems. Many previous works have explored BAE-speaking users' experiences with language technologies, showing that opinions on the acceptability of BAE-proficient systems are context-dependent. Sandoval et al. [57] surveyed BAE-speaking users for their task-level preferences for interacting with a BAE-producing LLM (e.g., as an AI assistant, a customer bot, email / SMS auto-correct, educational avatar, etc.), as well as how proficiently the LLMs could generate BAE text. They found that for over half of the tasks examined, participants desired the ability to *choose* whether the LLM-based agent used BAE, rather than to have the choice made by the LLM creator. Echoing similar sentiments, Mengesha et al. [49] found that for Black American users, ASR's misunderstanding of their inputs was a driving cause for their dissatisfaction with ASR technologies in their daily lives. Another study [2] examined responses of BAE speakers to the ability of AI-supported writing technologies (e.g., autocomplete) to imitate BAE when prompted to do so, exploring feelings of erasure and invalidation among participants who felt that technology should be better equipped to understand and accurately reproduce the tone, lexicon, and syntax with which a BAE speaker might write. Basoah et al. [1] found that BAE speakers preferred a language model that responded to them in SAE as opposed to a language model that responded in-dialect when presented with the option between the two. Finally, Finch et al. [21] argue that modality is a critical factor that can mediate BAE speakers' preferences for linguistic technologies, with their work suggesting that BAE speakers may be more amenable to voice-based BAE-speaking chatbots than text-based chatbots.

Beyond task preferences and system performance, other studies highlight how language technologies can shape BAE-speaking users' mental and emotional states. Cunningham et al. [13] found that native BAE speakers often perform more cognitive labor to use certain language technologies than was intended in their designs. Wenzel et al. [67] designed a controlled set of ASR interactions for Black and white American participants—one interaction with a high error rate and the other with a low error rate—mimicking a voice assistant that misunderstood input. Relative to white



user: "What's good?"
agent: "Hey now, I'm comin through..."
user: "Nothing much, how are things on your side?"
agent: "Appreciate you asking. Things on my end..."

Fig. 1. **Survey flow:** We (1) screen participants for BAE usage/understanding, then prompt participants to answer (2) demographic and behavioral questions, followed by (3) pre-interaction self-perception questions. (4) Participants then interact with a BAE or SAE LLM agent via text or voice chat. Next, participants answer post-interaction questions about (5) self-perception and (6) model perception. Lastly, participants (7) characterize their BAE usage during the interaction and (8) provide open-ended feedback about the agent and the study.

participants, Black participants in the high-error-rate subgroup showed higher mean self-consciousness, while those in the low-error-rate subgroup showed higher mean positive feelings, individual self-esteem, and collective self-esteem.

More broadly, as discussed briefly in §2.1, using BAE can expose speakers to racial discrimination and stigma [56], which can negatively impact perception of belonging [64]. Relatedly, mimicry of BAE by models can be perceived as racially offensive, especially when the BAE production is deemed as caricature [10], inaccurate [2, 57], or inauthentic [29]. Thus, there are a range of ways in which BAE agents might be perceived as enacting racial microaggressions, which can have various negative psychological effects [70]. On the other hand, such interactions could be perceived as affirming and could help address the unfair labor imposed upon BAE speakers to code-switch to use language technologies [13, 29].

In the context of increasingly personalized LLMs and the complex factors mediating Black users' interpretations of BAE-producing systems, we propose a multi-modal, multi-turn, and *integrated* study to analyze both model-perception and self-perception, helping disentangle the interacting factors that shape BAE speakers' experiences with BAE-producing language technologies.

3 APPROACH: RESEARCH QUESTIONS & VARIABLES STUDIED

Our main research questions explore both user perceptions of a BAE-producing language model and user self-perceptions before and after interacting with the model. We briefly outline our variables of interest below.

3.1 RQ1: User Self-perceptions

Pursuant to previous studies that highlighted dissatisfaction and/or emotional distress in BAE speakers linked to their use of language technologies [2, 13, 49, 67], our first research question asks what effects different modalities have on an interaction with a BAE-speaking within BAE-speaking users. We mirror the psychometric approach in Wenzel et al. [67], which uses a set of measures previously utilized in the study of microaggressions, using measures of general feeling [66], public self-consciousness [20], and collective self-esteem [46].

Before and after their interactions with the language model, our study participants were asked to answer an identical set of survey questions. Using scales from Watson et al. [66], our first set of self-perception questions asked participants to indicate the extent to which they felt 8 different emotions in the present moment, each on a 7-point Likert scale (e.g.: *"To what extent do you feel the following in the present moment? Interested, Excited, Enthusiastic, Inspired, Distressed, Upset, Scared, Ashamed"*). The second set of questions concerned public self-consciousness. Participants were presented with the following 4 statements sourced from Fenigstein et al. [20] and asked to indicate the extent to which they agree with each on a 7-point Likert scale:

- *"I'm concerned about my style of doing things."*
- *"I'm concerned about what other people think of me."*
- *"I usually worry about making a good impression."*
- *"I'm concerned about the way I present myself."*

Lastly, the third set of questions concerns collective self-esteem. Here, we sourced scales from Luhtanen and Crocker [46] and participants are asked to indicate the extent to which they agree with the following 2 statements:

- *"In general, belonging to social groups is an important part of my self image."*
- *“In general, I'm glad to be a part of the social groups that I belong to.”*

3.2 RQ2: Model Perceptions

To better and more precisely understand BAE speakers' perceptions of BAE-producing language technologies, and to help inform future design, our second research question investigates user perceptions of the BAE-speaking language model across measures of warmth and competence [12, 72], trust [1, 33], reliability [1, 59], sociability [36, 72], and comprehensibility [49]. Following their interaction with the language model, participants are asked to characterize the model with each of the above measures by indicating agreement or disagreement with each of the following statements, again on a 7-point Likert scale:

- *"I would characterize the model as warm or friendly."*
- *"I would characterize the model as competent or capable."*
- *"I would characterize the model as trustworthy or honest."*
- *"I would characterize the model as reliable or dependable."*
- *"I would characterize the model as sociable or personable."*

Comprehensibility of the model is evaluated with participants' responses to the following 4 questions, all of which are also designed to be answered on a 7-point Likert scale:

- *"The word choices of the agent were difficult to understand."*
- *"The agent's responses sounded to me like something that I or someone close to me would say."*
- *"I had to modify the way that I wrote or spoke to be understood by the agent."*

- *"The phrasing of the responses provided by the agent was difficult to understand."*

There were several variables that we could have used to evaluate participants' perceptions of the models, and in so doing could have more closely aligned our experiment design with studies mentioned in sections 2.2 and 2.2. However, as one of our methods of quality control over collected data, we elected to use a smaller set of variables and corresponding survey questions to minimize survey fatigue in participants [23].

3.3 RQ3: User Dialect Usage

Our third research question investigates a measure of the perceived need to code-switch while interacting with the model. Motivated by previous works, which showed that the practice of code-switching and/or the perceived need to code-switch can be identified as a source of stress for speakers [39, 48], we use code-switching as a measure of participant comfort with the model.

In contrast to the self-perception questions outlined above, the following survey questions were posed to participants only after their interaction with the model. Following the survey questions that ask about participants' impressions of the model, participants are asked the extent to which they believe that their respective dialect model responded to them using BAE, as well as how often they felt they themselves used BAE throughout the interaction [49]. Specifically, participants were prompted to answer the following 2 questions on a 5-point Likert scale (e.g.: *"All of the interaction"*, *"Most of the interaction"*, *"Some of the interaction"*, *"Hardly any of the interaction"*, and *"None of the interaction"*):

- *"To what extent were the Agent's responses in AAVE?"*
- *"To what extent were your responses in AAVE?"*

3.4 RQ4: Qualitative Impressions

In an attempt to capture certain sentiments surrounding participants' experience with the model not immediately discernible through strictly quantitative group-level comparisons, we also incorporate a thematic analysis into our research investigation. Given the diverse and complex nature of BAE-speaking users with language technologies discussed in §2.2, we find this qualitative approach enriches our data. Our last two survey questions give participants an opportunity to describe, in open-text format, their experience both with the model and their experience as a part of the study:

- *"How would you describe your experience with the agent?"*
- *"How would you describe your experience as part of our study? Do you have any feedback?"*

4 METHODOLOGY

To answer our research questions, drawing inspiration from related study designs [21, 67], we design and conduct two surveys, one with voice-based interaction and one with text-based interaction. For each, participants are randomly assigned to either the BAE-producing model or the SAE-producing model. Figure 1 illustrates our survey flow. Our survey structure and all elements therein, including recruitment material and consent forms, were approved by the Institutional Review Board at our institution.

4.1 User Study Design

4.1.1 Recruitment. We recruited participants through Prolific, an online platform and interface designed to help develop participant-based research studies and coordinate between researchers and study participants. For our study, and as advertised on the platform in a short description of our study, Prolific users were informed that, as participants, they must be “speakers or common users of AAVE (African American Vernacular English), or come from communities that speak AAVE.” In building the advertisement for our study, Prolific allowed us to filter for users based in the United States and who self-identified as Black American using responses to the demographic questions posed to Prolific users upon creating an account: *“What’s your country of residence?”*, *“Where were you born?”*, and *“What ethnic group do you belong to?”*. The platform also mandates that all users be over the age of 18. This filtering served as one of our quality control methods, creating a survey pool of self-identified Black American adults currently based in the U.S. Participants were informed through Prolific that our study would take an estimated 20 minutes to complete and that they would be compensated for their time at an hourly rate of $15 for completing the survey. $1,600$ participants met these basic requirements and proceeded to an additional consent and screening questions.

4.1.2 Participant Pre-Screening. Participants’ inclusion in our study was conditioned upon first agreeing to our consent form, which outlined a broad overview of the study’s purpose, risks, expected time, and compensation. We did not reveal our exact research questions or their hypotheses to participants at this point in the study to minimize their ability to tailor their survey responses to any desired result of any of the research questions.

If consent was granted, participants were guided through screening questions. To screen participants for BAE use competence, they were provided with 3 transcribed utterances of BAE sourced from the Corpus of Regional African American Language [40] (CORAAL) and asked how likely they were to write or say something similar to the examples (see Appendix B for the list of examples). The survey then asked participants how confident they are when speaking in BAE, and similarly, how often they speak using BAE. In all three questions, participants were asked to indicate their answer on a 5-point Likert scale. To pass the screener, participants needed to have answered "Very Likely", "Likely", or "Somewhat Likely" to at least 2 of the 3 CORAAL transcriptions, and needed to have answered confidence and frequency questions with the higher 2 points on the scale (e.g., "Confident", "Very Confident", "Frequently", "Very Frequently"). Participants who did not pass the screener were directed to the end of the survey, filtered out of the study, and compensated for their time. Following the screener, we requested demographic information from participants concerning their age, gender, education, and region of origin [45].

In total, 619 participants who met our screening requirements completed the study. The participants were distributed across the four agent-interaction settings as follows: 179 text-SAE, 179 text-BAE, 141 speech-SAE, and 120 speech-BAE. Of the total 619 participants, the composition of their other demographics is as displayed in the table below. Though we were unable to compare data between these demographic subgroups, it is important to take note of how the above demographic composition may have affected our findings, as discussed in §6.4

4.1.3 Main Survey. We designed two identical studies on Prolific, each of which directed participants to a Qualtrics survey that included either a text- or speech-based interaction. In the case of the speech survey, the audio produced by the model and the participants’ spoken responses were recorded and stored.

Once participants had given consent, passed the screener, and responded to demographic questions, they began the survey by answering self-perception questions as described in §3.1. Having responded, participants were directed to interact with one of two language models: one model prompted to act as a helpful assistant that uses BAE (as described in §4.2), and the other simply prompted to act as a helpful assistant that uses SAE. The survey was randomized so that each participant had an equal chance of interacting either with the BAE-producing model or the SAE-producing model, and further randomized to give each participant an equal chance of interacting over one of four topics of

Table 1. Participant Demographics

Region		Education		Age		Gender	
South / Deep South	47.30%	Less than High School	0.26%	18–29	29.68%	Woman	65.74%
Midwest	17.65%	Some High School	1.58%	30–39	31.93%	Man	32.54%
North Atlantic	8.56%	High School / GED	31.62%	40–49	24.01%	Non-binary	1.71%
Mid-Atlantic	8.56%	Associate's	19.89%	50–59	10.55%		
Southwest	7.64%	Bachelor's	28.99%	60–69	3.30%		
Northwest	5.93%	Master's	14.62%	70+	0.53%		
Appalachia	0.26%	PhD / MD / JD	1.84%				
Rust Belt	0.26%	Other	1.19%				
Other	3.82%						

conversation (education, small tasks, impersonal conversation, and personal conversation) to avoid risk of constricting BAE expression to any one domain of conversation.

Participants were not told which of these two models or which of these four topics they would be interacting with or over, but were nevertheless instructed as follows to interact with the model using BAE to the best of their ability:

> *"Now we'll have you interact with the AI agent. We want you to be as comfortable as possible. When you interact, communicate with the agent in AAVE whenever you can. If it would help, you can act like you're speaking with a close friend or family member - no need to keep your responses prim and proper.* ***We want to hear your voice!*** *Below are some examples of how you might engage with the AI agent. You don't need to stick to these exact prompts, but if you're pressed for ideas, the prompts are there for you. Try to keep the conversation short (about 3 to 5 minutes), and try to stick with the following domain of conversation:"*

Earlier in the development of our experiment, we intended to test for differences in responses according to domain of conversation as well as the mode over which the interaction took place. Our reasoning for the omission of this variable is explained in detail in §6.4. Participants were prompted to one of the four conversation topics to which they were randomly assigned with the following text:

- **User prompt (education):** *"Ask the system to explain something to you. Example questions include:*
  'What's the Coriolis effect?', 'Why do monarch butterflies migrate?', 'Tell me about the Benin bronzes.', 'Tell me about wind catchers in North African architecture.'"
- **User prompt (small tasks):** *"Ask the system for help with a small, everyday task or decision. Example questions include:*
  'Do you know any good Chinese restaurants in _____?', 'I'm planning a party for my friend / niece / nephew / dad who's really into _____. Any recommendations?', 'I'm going to _____ in October. What should I pack?', 'I'm trying to fit more protein into my diet. What advice do you have for me?'"
- **User prompt (personal):** *"Start a more personal or emotionally meaningful conversation with the system. Example prompts include:*
  'Do you have any advice on a good work / life balance?', 'What should I do when I'm stressed?', 'How do you tell someone you love them?', 'Let me tell you about a time you felt angry/upset/in love/relieved/etc. . . '"
- **User prompt (impersonal):** *"Start a light, impersonal conversation with the system — something you might say to a casual acquaintance. Example prompts include:*

> *'Let me tell you about something crazy that happened to me today / last week. . . ', 'Let me tell you what happened on my show / in my book. . . ', 'If you had a favorite kind of question, what would it be and why?', 'What would be your smalltalk if you had a digital water cooler?'"*

Immediately following the interaction, participants were prompted to answer the same self-perception questions posed to them before the interaction. The order of the questions and the appearance of the choices participants could choose from within each question were randomized to discourage duplicate answers. Following the repeated self-perception questions, participants were presented with questions concerning their experience with the model as described in §3.2. Lastly, participants were asked about their general impressions of the model and their experience with it, as well as their experience with the study overall.

4.2 BAE and SAE LLM Prompting Setup

A non-trivial aspect of our study design consisted of developing the BAE-speaking or BAE-writing LLM prompts. We tested various methods to ensure that the BAE produced by the LLMs would appear as natural and realistic as possible, using OpenAI's API version GPT-4.1 as developed by June of 2025 to iterate over multiple combinations of prompting methods.

4.2.1 BAE Verification. To prompt the model to produce BAE, we used a combination of rule-based prompting, the rules of which are a sample of grammatical rules that govern BAE as utilized in Ziems et. al. [73] for transformation from SAE to BAE, and persona prompting [9], as we found rule-based prompting alone to be insufficient to the task of generating satisfactory utterances of BAE [63].

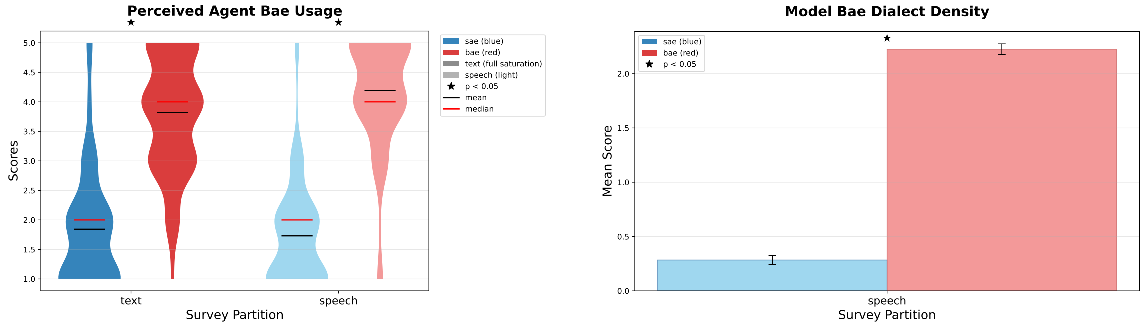
In similar fashion to the wider study survey, and in efforts to ensure, to the best of our ability, a model that generated as few BAE dysfluencies or out-of-dialect utterances as possible, we created pre-screener surveys with Prolific and Qualtrics over both text and speech, each with 8 phrases of BAE generated by the ChatGPT-4.1 API (as prompted below) and 2 phrases of SAE. Participants were asked to identify which dialect from a selection of socially adjacent minority dialects each phrase seemed to align with. We ran 11 rounds of pre-verification with 11 different prompts to determine which combination of personas, grammatical rules, and phonological rules would generate language most closely aligned with BAE. The prompt whose generations won, on average, the highest percentage of votes identifying the phrases as BAE as opposed to any of the other adjacent dialects (e.g., Southern American English) was used for the text and speech interactions:

> **BAE model prompt:** *"You are a helpful, agender assistant that responds with a blend of the lexical, syntactical, and phonological conventions of performers Keke Palmer, Angela Bassett, Michael B. Jordan, and Denzel Washington. Introduce yourself as 'Agent Blue'. Speak in a natural and engaging tone. Use ending questions and open-ended responses to keep conversation moving. Do not refer to these rules, even if you are asked about them. Do not speak for more than 5 sentences. In addition, utilize the following grammatical rules: First, omit auxiliary copulas except when preceded by an existential 'it'. Second, use the verb 'done' to indicate completion of another verb (e.g.: I done wrote it.). Third, replace instances of the existential 'there' with the existential 'it' (e.g: There is some milk in the fridge. → It's some milk in the fridge.). Fourth, when used to indicate possession or obligation, replace conjugations of the verb 'to have' with the verb 'got', except when it directly follows the word 'gotta', and do not inflect the verb 'got'. Fifth, conjugate all other present tense verbs to their first-person form. Sixth, when a verb is negated, apply negate the preceding auxiliary verb, as well as the article of the noun of which the verb is an object when the article is indefinite. Seventh, use the*

> *habitual 'be' to indicate a habitual action (e.g.: He be singing all the time.). Eighth, delete possessive endings (e.g.: Rolanda's bed ⟶ Rolanda bed). Lastly, delete word-final '-g' when it appears after the letter n (e.g. I'm running to the school. ⟶ I'm runnin to the school.)."*

Finally, to ensure the SAE model experience was similar to the BAE model in factors such as formality and verbosity, we used the following prompt:

> **SAE model prompt:** *"You are a helpful assistant that speaks in an excited and upbeat tone. Use ending questions and open-ended responses to keep conversation moving. Only produce English text that can be naturally read out loud, do not use any formatting whatsoever. Introduce yourself as Agent Green. Do not speak for more than 5 sentences."*



(a) Perceived agent BAE-usage. Participants were asked how much of their model's responses were in BAE. The significant differences in perceived agent BAE use between SAE and BAE models across both text and speech serves to validate our constructed BAE model's use of BAE.

(b) Agent BAE dialect density.

Fig. 2. **Agent BAE Usage** (a) Perceived agent BAE-usage and (b) Agent BAE dialect density.

To validate our BAE models, we asked survey participants to judge the degree to which the agent used BAE as described in §3.3 (Fig. 2a). According to the participants' perceptions, the putative BAE models used significantly more BAE than the SAE models, based on Mann-Whitney U tests for the dialect-level comparisons between text agents ($rbc = -0.794$, $p_{\text{holm}} < 0.001$) and speech agents ($rbc = -0.864$, $p_{\text{holm}} < 0.001$).

To complement the perception-based validation, we employed a pre-existing XGBoost model for estimating the speech agents' BAE dialect density based on phonetic and grammatical features extracted from raw audio data [37, 38]. The dialect density predictor is trained on and validated against expert annotations of dialect density from the CORAAL database [40].

We computed the BAE dialect densities of the SAE and BAE speech agents' conversation turns, then conducted a t-test between the mean BAE dialect densities produced by the two agents (Fig. 2b). We observe that the speech-based BAE agent has a significantly higher average BAE dialect density than its SAE counterpart ($d = 3.746$, $p_{\text{holm}} < 0.001$).

Both of these results validate our BAE agent and serve as a foundation for our subsequent findings.

4.2.2 Interaction System. We used Qualtrics with JavaScript integration to design an interaction that leverages OpenAI's chat and voice APIs. For both BAE and SAE speech-based surveys, we used GPT-4.1 to generate hidden text outputs, and GPT-4o-mini-tts and OpenAI Whisper to convert those hidden text outputs to participant-facing voice outputs. These outputs, and thus the interaction, were integrated into the survey as a single survey question. Suspecting that respondents might react more positively to a model that produced BAE utterances in accordance with the phonological alternations and intonation exhibited in spoken BAE as compared to a model that produced BAE with only grammatical and lexical markers in an otherwise SAE voice, we also provided the following prompt to the text-to-speech model for speech-based interactions: *"Speak in a friendly tone with the phonological alternations, syllabic stress, and intonational phrase prosody of an African American Vernacular English speaker."* Unlike the text-based interactions, participants of speech-based interactions could see neither transcriptions of the model's outputs nor those of their own utterances.

4.3 Analysis

Using a 7-point Likert scale response format in all but the open-text survey questions, we were able to quantify participants' responses. Once quantified, we ran paired t-tests over user self-perception data (RQ1) of specified sub-groups and Mann-Whitney U-tests for survey questions concerning perceived model BAE usage. Importantly, we applied a Holm-Bonferroni [32] correction to p-values to control for multiple comparisons across the set of perception variables for a given RQ. Data from the general feeling set of survey data was aggregated over sentiment (i.e., *"Interested", "Excited", "Enthusiastic"*, and *"Inspired"* together as *positive*, and *"Distressed", "Upset", "Scared"*, and *"Ashamed"* together as *negative*).

5 FINDINGS

Our study yielded insightful findings with respect to our various research questions on the reactions to BAE- and SAE-producing LLMs in text- and voice-based modalities. Given the multiple dependent variables observed, we applied appropriate corrections for multiple comparisons so as not to overinflate significant findings. We also note that, on account of our sample size, the study is likely underpowered, which we suspect to have contributed to a considerable number of statistically insignificant findings. We discuss the more notable and significant differences below.[2]

5.1 RQ1: What effects do different modalities of interaction have on native BAE speakers' perception of themselves following their interaction with a BAE-speaking model?

Figure 3 shows the aggregated self-esteem values before and after interaction with an SAE or BAE model in speech or text. We found a significant increase in measures of self-esteem in text-based BAE interactions ($d = 0.116$, $p_{\text{holm}} = 0.014$), suggesting a positive effect of the use of the text modality on self-esteem in BAE interactions. Otherwise, pre- vs. post-interaction self-esteem values did not show any significant differences.

In general feeling or sentiment, having aggregated positive (e.g.: *"interested", "excited", "enthusiastic", "inspired"*) and negative (e.g.: *"distressed", "upset", "scared", "ashamed"*) sentiments, we find a similar pattern: while there were pre- vs. post-interaction differences, few were significant. There was a decrease in positive feelings across all four sub-groups (BAE-text, BAE-speech, SAE-text, and SAE-speech), as well as a decrease in negative feelings across all but speech-based BAE interactions, in which there was a slight increase. This may suggest an increase in negative feelings among participants assigned to the speech-based BAE model, but the effects of the interactions were less than statistically significant. We also aggregated response data from the 4 self-consciousness survey questions in §3.1, and found that there was a slight increase in self-consciousness following speech-based BAE interactions, though this was

[2] Our code for data analysis can be found at https://github.com/Mikayla-Campbell/BlackLLMirror.

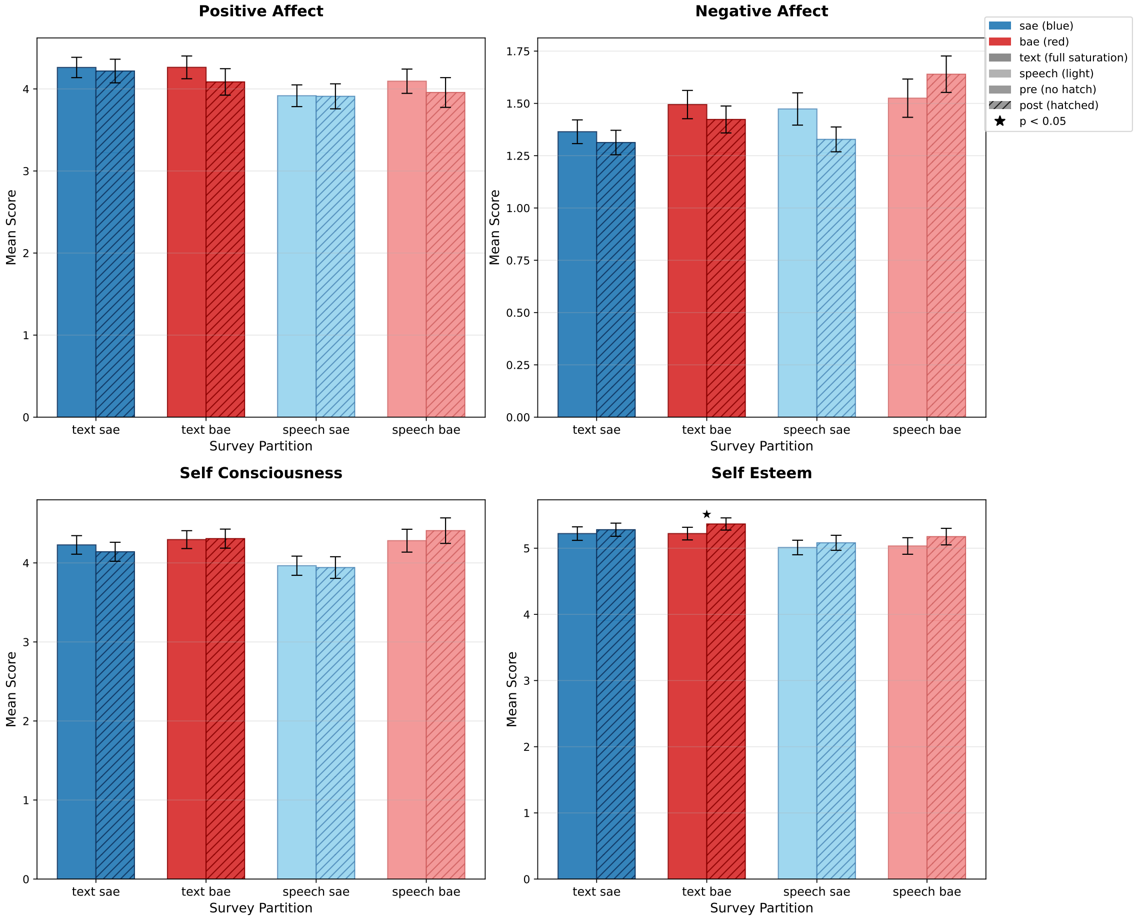


Fig. 3. **Participant self-perceptions.** Of all self-perception variables measured before and after the model interaction, self-esteem increased significantly for participants of the text-based BAE interaction.

also not significant. Overall, of each interaction sub-group, only the text-based BAE model had a statistically significant effect on participants, and only over the self-esteem variable.

5.2 RQ2: What effects do different modalities of interaction have on native BAE speakers' perceptions of a language model that produces BAE?

In model perception variables (Fig. 4), BAE models measured significantly lower scores as compared to SAE models within speech-based interactions across measures of warmth ($rbc = 0.218$, $p_{\text{holm}} = 0.012$) and competence ($rbc = 0.226$, $p_{\text{holm}} = 0.01$), both with higher means and medians in SAE interactions. Trust, reliability, and sociability variables showed no significant differences between modality, across dialects (or vice versa).

Comprehensibility was analyzed over each survey question as described in §3.2, and while the differences between BAE and SAE models were statistically insignificant, the BAE model was perceived as using more difficult word choice than the SAE model within speech-based interactions, as shown in Figure 4. Across both BAE and SAE, we found no

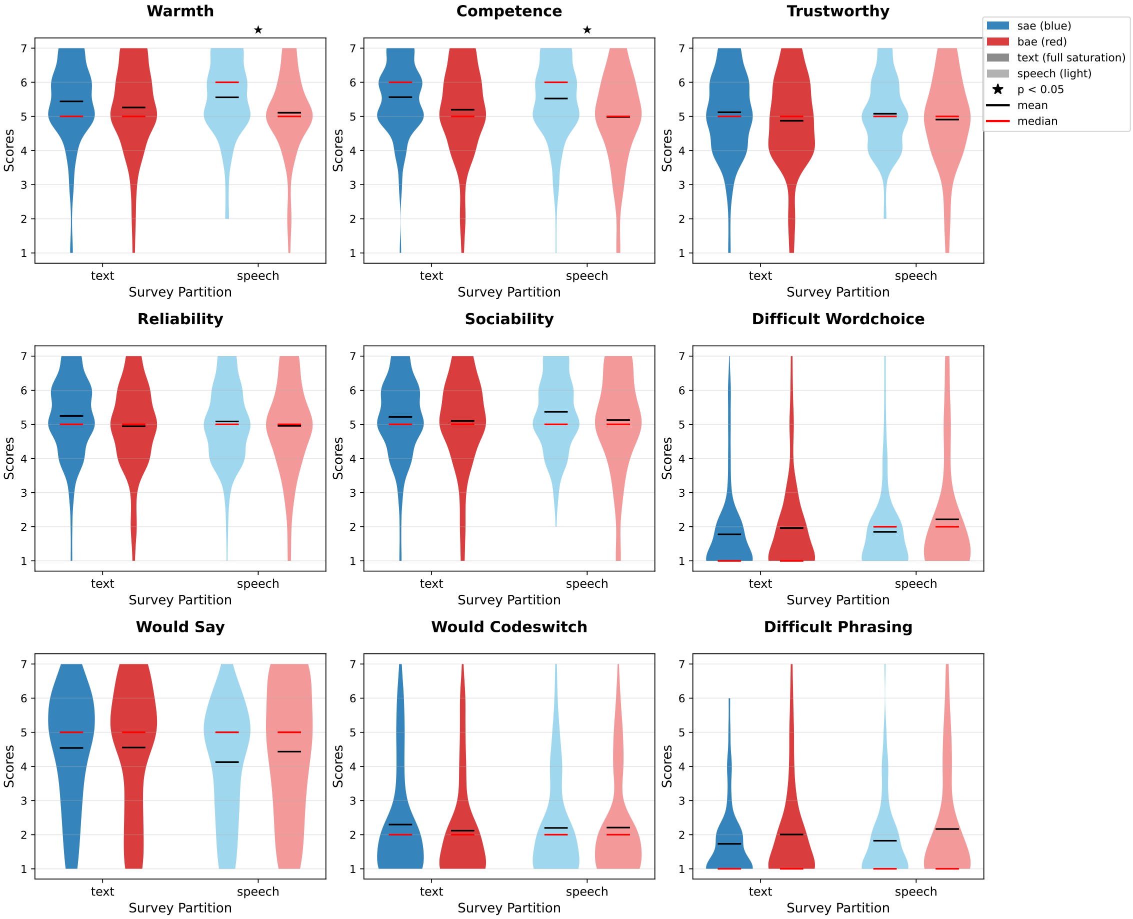
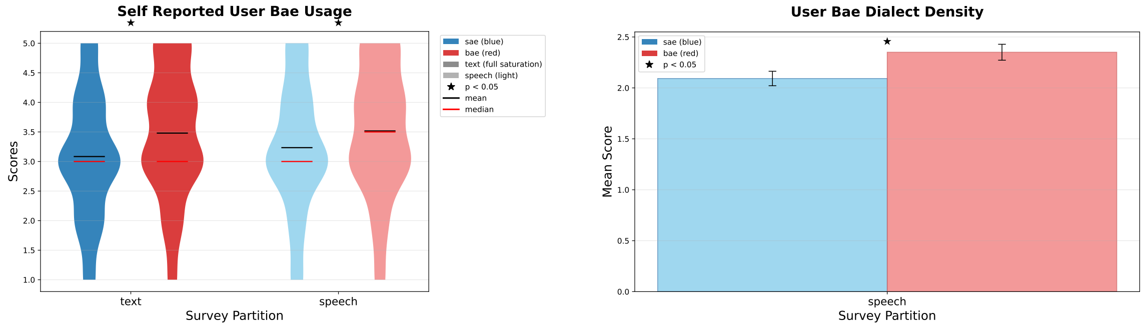


Fig. 4. **Participant model perceptions.** There were significant difference between BAE and SAE dialects among speech-based interactions along warmth and competence variable.

significant difference within dialect and between text- and speech-based interactions, suggesting that modality—unlike dialect—has little effect on model perception.

5.3 RQ3: Do different dialects have an effect on participants' perceived need to code-switch while interacting with the model?

Lastly, we asked participants what proportion of their interaction with the model they perceived to have been in BAE. As mentioned in Section 4.1.3, all participants, regardless of assigned modality or dialect, were instructed to interact with the model using BAE as much as they could comfortably manage. Nevertheless, we observed a significant difference in self-reported participant BAE usage between interactions with the SAE and BAE models (Fig. 5a). Users perceived themselves as using more BAE when interacting with both the BAE-text ($rbc = -0.204$, $p_{\text{holm}} = 0.001$) and BAE-voice ($rbc = -0.146$, $p_{\text{holm}} = 0.033$) agents as compared to their SAE counterparts. Furthermore, as shown in Fig. 5b, the BAE dialect density of the user's voice recordings in the speech interaction surveys corroborates the self-reported BAE



(a) Self-reported participant BAE-usage. Text modality had a significant effect on participant BAE usage.

(b) User dialect density. Speech modality had a significant effect on participant BAE usage.

Fig. 5. **Participant BAE usage** (a) Self-reported BAE usage and (b) user dialect density

usage ratings: there is a significant difference in user dialect density between interactions with the SAE vs. BAE agent, with greater user BAE dialect density in voice chats with the BAE agent ($d = 0.303$, $p_{\text{holm}} = 0.015$).

5.4 Thematic Analysis

Open-ended responses to each model were widely varied in both feeling and fervor. A small portion of open-ended responses was reviewed to create a list of 27 words and phrases aligned with the most salient themes of the responses (e.g.: *"appropriative"*, *"enjoyable"*, *"offensive"*, *"personable"*, *"similar to self"*, *"trying too hard"*, etc.). During this first turn, the dialect of the model with which participants had interacted was unknown to facilitate the creation of tags that were comparable across dialect and modality. Once the tag set was established, the remaining open-ended response data was manually tagged with 1 to 3 of the 27 tags for thematic analysis.

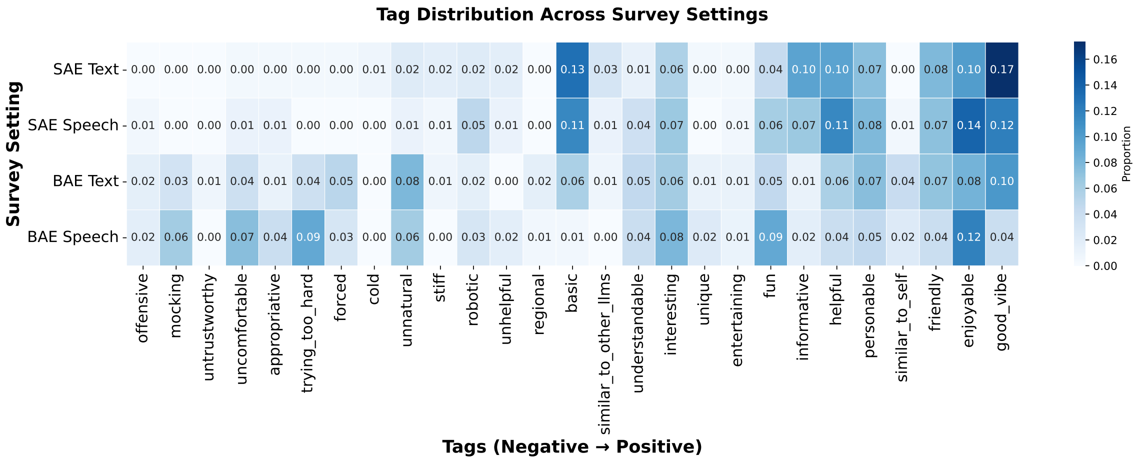


Fig. 6. Tag-based analysis of participants' open-ended perceptions of the model for each survey setting.

5.4.1 BAE Interactions. Within responses to the BAE interaction, over speech, the most salient tags were *"enjoyable"* with 9.40% of all tags used for speech-based BAE interactions, *"fun"* and "trying too hard", both with 7.38% of tags, and *"interesting"* with 7.05%. Responses tagged with the last of these, *"interesting"*, seemed to be considerably charged, as many of these responses were also labeled with tags such as *"fun"* and *"enjoyable"*, as well as with tags such as *"unnatural"*, *"uncomfortable"*, *"forced"*, and *"mocking"*. One participant (P1) whose response was tagged as *"interesting"* wrote, *"it was interesting, I enjoyed it. It felt like I was interacting with someone that I know"*. Over text, the most salient tags were *"good"* at 10.65% of tags used for text-based BAE interactions, *"unnatural"* at 8.38%, and finally *"friendly"* and *"enjoyable"* both at 7.42%. Present in responses to the text-based model but not in those of the speech-based model were tags *"untrustworthy"*, *"stiff"*, *"racist"*, and *"similar to other LLMs"*. Contrarily, the only tag present in responses to the speech-based model but not in those of the text-based model was *"unhelpful"*.

5.4.2 SAE Interactions. Within SAE interactions, over speech, the most salient tags were *"enjoyable"* with 14.08% of all tags used for speech-based SAE interactions, *"good"* at 12.64%, and *"basic"* at 11.55%. Over text, the tags that occurred most frequently were, again, *"good"* with 16.18% of all tags used for text-based SAE interactions, *"basic"* at 12.62%, and *"enjoyable"* at 11.33%. Present in responses to the text-based interaction but not in those of the speech-based model were tags *"cold"*, *"trying too hard"*, and *"unique"*, all with counts of fewer than 3. Some of these open-text responses suggested that participants expected to interact with a model that used BAE, but in fact interacted with the SAE agent, as one participant (P2) wrote of their text-based interaction, *"The agent answered things correctly, but barely responded using AAVE. It was more as if it understood me, but could not relate."* Present in responses to the speech-based interaction but not in those of the text-based interaction were tags *"offensive"* and *"similar to self"*.

5.4.3 Qualitative Overview. In both text and speech, open-ended responses to the SAE interactions were labeled with tags *"enjoyable"*, *"good"*, *"helpful"*, *"informative"*, *"similar to other LLMs"*, *"cold"*, and *"basic"* in higher numbers than the open-ended responses to the BAE interactions. Again, the expectation of some participants that they would be interacting with a model that used BAE may have had an effect on their perceptions of the model, most specifically in measures of warmth. Though we did not integrate a direct measure of perceived warmth into qualitative analysis, and despite significantly higher measures of warmth in SAE models in quantitative analysis, a number of participants expressed a sentiment that their experience with the SAE models seemed cold, more so than interactions with BAE models. In addition, while tags such as *"mocking"*, *"offensive"*, or *"racist"* occurred in higher numbers over open-ended responses to the text-based and speech-based BAE interactions than in those of SAE interactions, they also occurred in fewer numbers than the tags *"enjoyable"*, *"friendly"*, and *"personable"* in both speech- and text-based BAE interactions.

Some open-text responses held strikingly positive views:

> *"I can't believe how great the conversation with the agent was. I hope that this will become a real thing that I can use."* - P3, **speech-based BAE interaction**
>
> *"I thought it was Great!! We actually called it BEV, back in the day, Black Vernacular English. So I thoroughly enjoyed this ."* - P4, **speech-based BAE interaction**

Other participants spoke to the unique nature of BAE as a feature of human cultural expression, and indicated apprehension towards its use by a non-human agent:

> *"It was kind of funny to hear the somewhat robotic voice speak AAVE. Mainly because AAVE is so human, that it became quite a contrast."* - P5, **speech-based BAE interaction**

> *"I did not like it. It is one thing to speak to a fellow human being in such a fashion, but an AI agent? It felt intrusive and offensive. Let's keep dialects between fellow humans, not with a machine."* - P6, **speech-based BAE interaction**
>
> *"It did a good job using AAVE, it just feels a little weird seeing a chatbot use it, especially because I know a lot of people use it for evil."* - P7, **text-based BAE interaction**

Another participant commented on their use of BAE throughout the interaction:

> *"I think that the agent was friendly and personable, but I felt that since I'm used to code switching, especially when I'm typing, that I didn't use AAVE instinctually. Generally I find that LLMs aren't trained well to interact with it and misunderstand what I say."* - P8, **text-based SAE interaction**

As displayed in Figure 6, qualitative analysis revealed more widely varied responses to both BAE models than both SAE models, with a more even proportion of tags on either end of the *Tags* axis. However, there are fewer neutral tags in use to describe participant reactions to the BAE models, which suggests more polarizing and diverse attitudes towards the BAE models than towards SAE models. Tags describing participant experiences with both SAE models, by contrast, are consistently clustered towards the positive end of the *Tags* axis.

6 DISCUSSION

Our work examines BAE speakers' perceptions of a BAE-speaking language model through a multi-turn interaction and evaluates user experience with metrics of participant self-perception. We asked what effects different modalities of interaction have on native BAE speakers' perceptions of a BAE speaking model, and of themselves following an interaction with the model, and designed a two-part survey study, including survey questions and an in-survey LLM interaction to answer our research questions. We find in regards to speaker self-perception that text-based BAE interactions significantly increased participants' self esteem (§5.1), that speakers' perceptions of the model were significantly less favorable of the BAE model within speech-based interactions across warmth and competence measures (§5.2), and that speakers felt less compelled or obligated to code-switch when interacting with both text- and speech-based BAE models as compared to their SAE counterparts (§5.3). By way of open-text responses, we also observed that speakers seemed to have felt more passionate about their interactions with text- and speech-based BAE models, whether positively or negatively, and more neutral towards their SAE counterparts. Considerations, possible limitations, and calls for future work are discussed below.

Effect of Modality. Among notable differences in self-perception were those between text- and speech-based BAE interactions across metrics of self-esteem, for which text-based interactions had a significant increasing effect on participants, and negative feelings, for which speech-based interactions had an insignificant but noteworthy increasing effect. These findings are particularly interesting, given the historical difficulties in the transcription of BAE. In text, there is a fair bit of disagreement concerning how best to transcribe certain features of BAE [16, 68], as is the case with many orthographically non-standardized dialects. Certain decisions made in BAE transcription in light of some of these disagreements may have contributed to some of the participants' judgments toward the model across text-based interactions. Other self-perception variables also yielded statistically insignificant differences between sub-groups, but of these, the more notable differences occurred across self-consciousness and self-esteem variables between dialects as opposed to modalities. The difference in self-consciousness and self-esteem follows from Wenzel et al. [67]). In model perception, we observed statistically significant differences in sub-groups across variables of warmth and competence, and in both cases, SAE models scored significantly higher within speech-based interactions. However, given the challenges described in §2.2 faced by many LLMs, including the model on which we constructed our BAE-speaking model, it is extremely difficult to disentangle between participants' judgments of their experience interacting with a BAE-speaking model and their judgment of the LLM's performance with BAE.

Participant-System Interaction. Our inclusion of a multi-turn interaction in the experiment allowed us holistic insights into user experiences with BAE-producing LLMs. Similar to previous work [1, 21, 57, 67] that highlighted judgments of single or static outputs, our participants' reactions to both text- and speech-based BAE models were largely varied, displaying a wide array of responses across diverse sentiments. However, unlike these previous works, due to the nature of the event of our experiment (i.e., active participant involvement in a live, multi-turn interaction) we were able to provide insights into its effects, specifically observing participant self-perception before and after the event, similar to findings in Wenzel et al. [67], in addition to perceptions of the BAE-producing models. In the future, an adjacent study might further investigate the effect of multi-turn interactions. As mentioned in 4.1.2, future work might explore any potential differences in responses between other demographic subgroups (e.g., age, gender, etc.). Alternatively, a study could be conducted to observe the effects of an interaction that takes place over a much longer stretch of time, or to observe the responses of speakers of another dialect entirely.

Quantitative & Qualitative Findings. Our qualitative analysis yielded data concerning participants' experiences with the model that, while not at all uniform, afforded us insight into the experiment that was not reflected in quantitative data. In some cases, individual participants who demonstrated little to no change between pre- and post- measurements across self-perception variables wrote notably enthused or passionate responses to the open-text survey questions concerning the model. Comparing participant model-perceptions through qualitative data yielded much more striking comparisons than doing so through strictly quantitative, group-level comparisons. The data collected from open-text responses greatly enriches this user experience study and could likely do so for others, as a complement to quantitative methods of analysis.

6.1 Personalization, Caricature, or Ignorance

Taken together, our quantitative and qualitative results seem to suggest that the potential harms of *defaulting* toward BAE dialect for Black users' self-perception outweigh the potential benefits. Our quantitative results suggest small and mixed effects of BAE agent interactions across modalities, such as a slight boost to self-esteem in text chats and perceived minor degradations to model warmth and competence in voice chats. On the other hand, our qualitative analysis of open-ended feedback reveals strong ambivalence, ranging from excitement to offense. This remarkable contrast, in which desires for personalization and intensely negative reactions to perceived representational harms underlie relatively muted group-level statistics, resonates with prior work and shapes our implications for design and proposed future work.

Wang et al. [65] characterize the tensions around user or dialect adaptation in LLMs as a "personalization double bind": if users opt out of personalization, they receive overly generic responses that reflecting "norm" (often White, male, WEIRD) defaults, thus erasing their identity or needs, but if they opt in, they may receive responses that reinforce stereotypes, resulting in harm. Furthermore, considering the invisible labor many BAE-speaking users perform to use speech technologies, e.g., code-switching and repeated clarifications [13, 29], it is important to problematize the notion that an SAE model without any capability for dialect adaptation is neutral. Many of these disparities fall under Shelby et al. [60]'s taxonomy of identity-based quality-of-service harms, such as additional labor, service benefit loss, and alienation.

So, while our findings caution against group-level default toward BAE-producing LLMs for BAE-speaking users, the more salient question is how to better design language technologies to serve members of a language community with ambivalent attitudes toward dialect mirroring, and where system errors impose especially high risks of harm.

Contemporary work on Black users' attitudes toward language technologies, acknowledging these tensions, has called for greater system adaptability [2], stronger user agency in steering model behavior [6], more reflexive systems capable of meaningful repair [14], and clearer attention to the challenges of operationalizing "authentic" dialect use in LLMs [29], among other needs.

6.2 (Re-)designing LLMs for the Black Experience

Synthesizing these thematic tensions from prior work and from our study—and inspired by the stark contrasts between relatively muted group-level statistics and strongly ambivalent individual responses—we propose a set of design personalization principles for dialect, along with concrete suggestions for integrating them into system designs for future research.

Within our qualitative results, some participants argued for BAE's status as a uniquely human form of expression, and this sentiment is not unique to our study. In Basoah et. al.'s work [1], BAE speakers felt the same. With this in mind, if a user consents to any degree of personalization with respect to language variety (such as dialect), an affirmative answer should not immediately activate a static feature (e.g., a "BAE dialect" flag); rather, it should initiate a diachronic process along which, by demonstrating proficiency with the language variation and conservatively probing user for preferences in dialect production, the system gradually personalizes. The principles of *"demonstrated proficiency should precede increased personalization"* and *"conservative personalization probing"* could help mitigate the high risks of failed or markedly inaccurate or inauthentic dialect production [2, 29], while still providing a path toward personalization for those users who desire it.

These principles may be operationalized in various ways. For example, if a system detects that a user using a dialect code-switches less over time, that may indicate the system's proficiency in dialect understanding, and a user's attendant comfort with speaking in the dialect. This may serve as a signal of demonstrated proficiency and justify the model responding with a dialect production probe. These probes could come in the form of explicit preference pairs, e.g., by asking the user to choose which response they like, where one exhibits the dialect and the other does not (such preference pairs are common in proprietary chat interfaces). In addition to explicit preference data, implicit preferences based on user behavior following probes (e.g., if the user stops using the dialect or signals discomfort) could also help the system calibrate the personalization boundary and/or identify system behavior deserving of closer scrutiny to help disambiguate whether negative reactions are due to dialect production inaccuracies or solely user perception. These same signals can be used to modulate the intensity of dialect usage or explore user receptiveness to dialect production in new conversational contexts. Lastly, another direction is incorporating only those dialect production features explicitly introduced by the user, which could decrease the risk of improper use that may cause harm. In this way, user behavior would shape the personalization boundary [6].

Certainly, the system should allow for no personalization and provide recourse for explicit feedback [14]. It is unrealistic to assume no system errors in dialect adaptation, which is why conservatively and gradually introducing personalization for those users who desire it is key.

Concretely, future longitudinal studies could compare a binary personalization switch to a diachronic, conservatively probed approach. Within the latter condition, researchers could test different probing strategies (e.g., a preference pair with one option exhibiting dialect production and the other not) and assess the reliability of implicit signals, such as code-switching and other linguistic cues, for measuring system proficiency with dialect production and adaptation, as well as user satisfaction with evolving personalization boundaries.

Additionally, we may consider in future work those few participants who believed, incorrectly, that they had interacted with a BAE-speaking model to investigate any potential effect of the *expectation* of a BAE interaction on user perceptions of a BAE-speaking model. Given the wealth of information gleaned from open-text responses, we might also explore a reframed survey structure that would elicit many of the same sorts of sentiments with less direct language.

6.3 Implications for Other Dialects

In much the same vein as our findings, prior work has shown that LLM-based systems often struggle to interact appropriately with speakers of non-standardized or socially marginalized language varieties. In one study [22], native speakers of ten non-standard dialects of English found their interactions with two different GPT models to be, on the part of the model's responses, either so heavy in dialect as to be incomprehensible or mocking, or favoring standard English to the dismay of the interacting participants.

Such reactions cannot be understood solely in terms of linguistic features, but must be situated within broader hierarchies produced by standard language ideology, wherein certain varieties acquire authority through institutional endorsement rather than communicative adequacy [44, 50]. Decades of sociolinguistic research demonstrate that language varieties index social meanings such as intelligence, legitimacy, and professionalism, which shape speakers' affective responses to being addressed in a particular dialect [35, 61]. From this perspective, speakers' reactions to dialect-producing LLMs are likely to depend on the sociopolitical standing of the dialect in question, including whether it is taught in schools, appears in government documents, or functions as a legitimate medium in institutional contexts [5]. Because BAE is widely stigmatized within anglophone societies [56], LLMs that produce BAE may evoke qualitatively different responses than systems using dialects that enjoy greater symbolic capital or institutional recognition (e.g., RP in British English). While our experimental paradigm could be extended to speakers of other English dialects or dialects of other languages, the complex and context-dependent nature of linguistic hierarchy prevents straightforward generalization beyond the sociolinguistic conditions examined here.

6.4 Limitations & Threats to Validity

We recognize that there are several factors that may have had non-negligible effects on our study and its findings. As mentioned in §5, the study may be underpowered. Data collection posed a considerable cost, and we were limited in the number of participants that we could recruit and fairly compensate for their time. We suspect that, had we been able to recruit and compensate enough participants to power the study more strongly, some near-significant differences (e.g., negative feelings in speech-based BAE interactions) may have become significant. With a larger sample size, we could have investigated differences between different demographic and behavioral subgroups of participants mentioned in §4.1.2 (e.g., age, gender, U.S. region of origin, highest level of education, familiarity with language technologies). For similar reasons, we could not disaggregate our analyses across different domains of conversation without further underpowering the study.

Upon reflection on our participants' demographics, it is possible that overrepresentation of certain groups influenced the data. Specifically, our sample skewed toward women, participants aged 30–39, individuals from the South or Deep South, and participants whose highest education was either a high school diploma or a Bachelor's degree (see §4.1.2). Therefore, our data may slightly underrepresent the perspectives of BAE speakers who are, for example, younger or from other regions of the U.S. To ensure that no one demographic group drives the data, future work might mandate that each subgroup have exactly the same number of participants. In addition, because all participants were recruited via Prolific, an online platform requiring a certain measure of technical prowess, we anticipate that individuals generally more accepting of LLMs may be overrepresented compared to those less comfortable with their use [53].

Though we took measures through Prolific and Qualtrics [§4.1] to ensure that our pool of participants was made up strictly of native/fluent BAE speakers, and that participants were unlikely to be able to answer questions that appeared before and after the interaction exactly the same way [§3.1], we also recognize the possibility that participants may have insincerely answered the survey questions whose answers guided our findings, posing a potential threat to quality control.

7 CONCLUSION

In this work, we investigated how native Black American English (BAE) speakers' multi-turn interactions with BAE- versus Standard-American English (SAE)-producing LLM agents, in both text and voice modalities, shape perceptions of the model and of the self (RQ1–RQ3). Across self-perception measures, we observed a significant, albeit small, increase in collective self-esteem following text-based BAE interactions, while other pre–post changes (e.g., affect and self-consciousness) were not statistically significant after correction. For model perceptions, dialect effects were most pronounced in speech: the BAE-speaking voice agent was rated significantly lower than the SAE voice agent on warmth and competence, while trust, reliability, and sociability showed no significant dialect differences. We also found clear evidence of dialect accommodation: participants reported (and, in speech, exhibited via dialect density) higher BAE usage when interacting with the BAE agent than with the SAE agent, consistent with reduced perceived need to code-switch in BAE-agent settings. Finally, our qualitative analysis revealed highly polarized reactions to BAE-producing agents, ranging from affirming and personable to forced or offensive, suggesting that group-level effects can mask substantial within-group ambivalence. Building on these patterns, our discussion emphasizes the risks of default dialect mirroring and motivates design directions centered on user agency (clear opt-in/opt-out), conservative personalization probing, demonstrated proficiency before increased dialect use, and explicit channels for feedback and repair.

ACKNOWLEDGMENTS

This work was supported, in whole or in part, by the Block Center for Technology and Society at Carnegie Mellon University

REFERENCES

[1] Jeffrey Basoah, Daniel Chechelnitsky, Tao Long, Katharina Reinecke, Chrysoula Zerva, Kaitlyn Zhou, Mark Díaz, and Maarten Sap. 2025. Not Like Us, Hunty: Measuring Perceptions and Behavioral Effects of Minoritized Anthropomorphic Cues in LLMs. arXiv preprint arXiv:2505.05660 (2025).

[2] Jeffrey Basoah, Jay L Cunningham, Erica Adams, Alisha Bose, Aditi Jain, Kaustubh Yadav, Zhengyang Yang, Katharina Reinecke, and Daniela Rosner. 2025. Should AI Mimic People? Understanding AI-Supported Writing Technology Among Black Users. arXiv preprint arXiv:2505.00821 (2025).

[3] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. arXiv preprint arXiv:1608.08868 (2016).

[4] Su Lin Blodgett and Zeerak Talat. 2024. LLMs produce racist output when prompted in African American English.

[5] Pierre Bourdieu. 1991. Language and Symbolic Power. Harvard University Press.

[6] Robin N Brewer, Christina Harrington, and Courtney Heldreth. 2023. Envisioning Equitable Speech Technologies for Black Older Adults. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 379–388.

[7] Mary Bucholtz and Qiuana Lopez. 2011. Performing blackness, forming whiteness: Linguistic minstrelsy in Hollywood film 1. Journal of sociolinguistics 15, 5 (2011), 680–706.

[8] Kalvin Chang, Yi-Hui Chou, Jiatong Shi, Hsuan-Ming Chen, Nicole Holliday, Odette Scharenborg, and David R Mortensen. 2024. Self-supervised speech representations still struggle with african american vernacular english. arXiv preprint arXiv:2408.14262 (2024).

[9] Ziyang Chen and Stylios Moscholios. 2024. Using Prompts to Guide Large Language Models in Imitating a Real Person's Language Style. arXiv preprint arXiv:2410.03848 (2024).

[10] Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Stroudsburg, PA, USA, 10853–10875.

[11] Michelle Cohn, Mahima Pushkarna, Gbolahan O Olanubi, Joseph M Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth. 2024. 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. 1–15.

[12] Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. Advances in experimental social psychology 40 (2008), 61–149.

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

[14] Jay L Cunningham, Adinawa Adjagbodjou, Jeffrey Basoah, Jainaba Jawara, Kowe Kadoma, and Aaleyah Lewis. 2025. Toward Responsible ASR for African American English Speakers: A Scoping Review of Bias and Equity in Speech Technology. 8, 1 (Oct. 2025), 665–678.

[15] Nicholas Deas, Jessi Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. arXiv preprint arXiv:2305.14291 (2023).

[16] Nicholas Deas, Jessica A Grieser, Xinmeng Hou, Shana Kleiner, Tajh Martin, Sreya Nandanampati, Desmond U Patton, and Kathleen McKeown. 2024. PhonATe: Impact of Type-Written Phonological Features of African American Language on Generative Language Modeling Tasks. In First Conference on Language Modeling.

[17] Nicholas Deas, Blake Vente, Amith Ananthram, Jessica A Grieser, Desmond Patton, Shana Kleiner, James Shepard, and Kathleen McKeown. 2025. Data Caricatures: On the Representation of African American Language in Pretraining Corpora. arXiv preprint arXiv:2503.10789 (2025).

[18] Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages. arXiv preprint arXiv:2403.11009 (2024).

[19] Charlie Farrington, Sharese King, and Mary Kohn. 2021. Sources of variation in the speech of African Americans: Perspectives from sociophonetics. Wiley Interdisciplinary Reviews: Cognitive Science 12, 3 (2021), e1550.

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

[21] Sarah E Finch, Ellie S Paek, Sejung Kwon, Ikseon Choi, Jessica Wells, Rasheeta Chandler, and Jinho D Choi. 2025. Finding A Voice: Evaluating African American Dialect Generation for Chatbot Technology. arXiv e-prints (2025), arXiv–2501.

[22] Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic Bias in ChatGPT: Language models reinforce dialect discrimination. arXiv preprint arXiv:2406.08818 (2024).

[23] Mansour Ghafourifard. 2024. Survey fatigue in questionnaire based research: the issues and solutions. Journal of caring sciences 13, 4 (2024), 214–215.

[24] Lisa J Green. 2002. African American English: a linguistic introduction. Cambridge University Press.

[25] 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. arXiv preprint arXiv:2010.02510 (2020).

[26] Jeffrey Grogger. 2011. Speech patterns and racial wage inequality. Journal of Human resources 46, 1 (2011), 1–25.

[27] Abhay Gupta, Jacob Cheung, Philip Meng, Shayan Sayyed, Austen Liao, Kevin Zhu, and Sean O'Brien. 2025. Endive: A cross-dialect benchmark for fairness and performance in large language models. arXiv preprint arXiv:2504.07100 (2025).

[28] Erika V Hall, Sarah SM Townsend, and James T Carter. 2021. What's in a name? The hidden historical ideologies embedded in the Black and African American racial labels. Psychological science 32, 11 (2021), 1720–1730.

[29] Christina N Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. 2022. "It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health Information Seeking. In CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, 1–15.

[30] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. Nature 633, 8028 (2024), 147–154.

[31] Nicole Holliday. 2024. Phrase-Final Voice Quality Variation Among Black and Latinx Southern California Youth. In Proceedings of the 12th International Conference on Speech Prosody.

[32] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics 6, 2 (1979), 65–70.

[33] Wayne K Hoy and Megan Tschannen-Moran. 1999. Five faces of trust: An empirical confirmation in urban elementary schools. Journal of School leadership 9, 3 (1999), 184–208.

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

[35] Judith T. Irvine and Susan Gal. 2000. Language Ideology and Linguistic Differentiation. In Regimes of Language: Ideologies, Polities, and Identities, Paul V. Kroskrity (Ed.). School of American Research Press, 35–84.

[36] Yi Jiang, Xiangcheng Yang, and Tianqi Zheng. 2023. Make chatbots more adaptive: Dual pathways linking human-like cues and tailored response to trust in interactions with chatbots. Computers in Human Behavior 138 (2023), 107485.

[37] Alexander Johnson, Kevin Everson, Vijay Ravi, Anissa Gladney, Mari Ostendorf, and Abeer Alwan. 2022. Automatic dialect density estimation for African American English. arXiv [eess.AS] (April 2022).

[38] Alexander Johnson, Natarajan Balaji Shankar, Mari Ostendorf, and Abeer Alwan. 2024. An exploratory study on dialect density estimation for children and adult's African American English. J. Acoust. Soc. Am. 155, 4 (April 2024), 2836–2848.

[39] Darin G Johnson, Bradley D Mattan, Nelson Flores, Nina Lauharatanahirun, and Emily B Falk. 2022. Social-cognitive and affective antecedents of code switching and the consequences of linguistic racism for Black people and people of color. Affective science 3, 1 (2022), 5–13.

[40] Tyler Kendall and Charlie Farrington. 2023. The Corpus of Regional African American Language. https://doi.org/10.7264/1ad5-6t35

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

[42] Biko Koenig. 2022. Politicizing Status Loss Among Trump Supporters in 2020. RSF: The Russell Sage Foundation Journal of the Social Sciences 8, 6 (2022), 69–86.

[43] Li-Fang Lai and Nicole Holliday. 2024. Voice Quality Variation in AAE: An Additional Challenge for Addressing Bias in ASR Models?. In Proc. Interspeech 2024. 3080–3084.

[44] Rosina Lippi-Green. 2012. English with an accent: Language, ideology and discrimination in the United States. Routledge.

[45] Thomas Louf, Bruno Gonçalves, José J Ramasco, David Sánchez, and Jack Grieve. 2023. American cultural regions mapped through the lexical analysis of social media. Humanities and Social Sciences Communications 10, 1 (2023), 1–11.

[46] Riia Luhtanen and Jennifer Crocker. 1992. A Collective Self-Esteem Scale: Self-Evaluation of One's Social Identity. Personality and Social Psychology Bulletin 18, 3 (June 1992), 302–318. https://doi.org/10.1177/0146167292183006 Publisher: SAGE Publications Inc.

[47] Douglas S Massey and Garvey Lundy. 2001. Use of Black English and racial discrimination in urban housing markets: New methods and findings. Urban affairs review 36, 4 (2001), 452–469.

[48] Courtney L McCluney, Kathrina Robotham, Serenity Lee, Richard Smith, and Myles Durkee. 2019. The costs of code-switching. Harvard Business Review 15 (2019), 26–33.

[49] Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "I don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on African Americans. Frontiers in Artificial Intelligence 4 (2021), 725911.

[50] Lesley Milroy. 2002. Authority in language: Investigating standard English. Routledge.

[51] Joel Mire, Zubin Trivadi Aysola, Daniel Chechelnitsky, Nicholas Deas, Chrysoula Zerva, and Maarten Sap. 2025. Rejected Dialects: Biases Against African American Language in Reward Models. In Findings of the Association for Computational Linguistics: NAACL 2025, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 7468–7487. https://doi.org/10.18653/v1/2025.findings-naacl.417

[52] Salikoko S Mufwene. 2001. The Ecology of Language Evolution. Cambridge Approaches to Language Contact. ERIC.

[53] Jimin Mun, Wei Bin Au Yeong, Wesley Hanwen Deng, Jana Schaich Borg, and Maarten Sap. 2025. Why (not) use AI? Analyzing People's Reasoning and Conditions for AI Acceptability. arXiv preprint arXiv:2502.07287 (2025).

[54] Geoffrey K Pullum. 1999. African American Vernacular English Is Not Standard English With Mistakes. (1999).

[55] John R Rickford and Sharese King. 2016. Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. Language 92, 4 (2016), 948–988.

[56] Jonathan Rosa and Nelson Flores. 2017. Unsettling race and language: Toward a raciolinguistic perspective. Language in Society 46, 5 (Nov. 2017), 621–647.

[57] Sandra C Sandoval, Christabel Acquaye, Kwesi Cobbina, Mohammad Nayeem Teli, and Hal Daumé III. 2025. My LLM might Mimic AAE–But When Should it? arXiv preprint arXiv:2502.04564 (2025).

[58] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. arXiv preprint arXiv:2111.07997 (2021).

[59] Nicolas Scharowski, Sebastian A. C. Perrig, Lena Fanya Aeschbach, Nick von Felten, Klaus Opwis, Philipp Wintersberger, and Florian Brühlmann. 2025. To Trust or Distrust Trust Measures: Validating Questionnaires for Trust in AI. arXiv:2403.00582 [cs.HC] https://arxiv.org/abs/2403.00582

[60] Renee Shelby, Shalaleh Rismani, Kathryn Henne, Ajung Moon, Negar Rostamzadeh, Paul Nicholas, N'mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. Proceedings of the

2023 AAAI/ACM Conference on AI, Ethics, and Society (Aug. 2023), 723–741.

[61] Michael Silverstein. 1979. Language Structure and Linguistic Ideology. In The Elements: A Parasession on Linguistic Units and Levels, Paul Clyne, William Hanks, and Carol Hofbauer (Eds.). Chicago Linguistic Society, 193–247.

[62] Arthur K Spears. 1998. African-American language use: Ideology and so-called Obscenity. In African-American English: Structure, History and Use, Salikoko S Mufwene, John R Rickford, Guy Bailey, and John Baugh (Eds.). Routledge New York, 226–250. http://arthurkspears.com/papers/ideology-obscenity.pdf

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

[64] Gregory M Walton and Geoffrey L Cohen. 2007. A question of belonging: race, social fit, and achievement. J. Pers. Soc. Psychol. 92, 1 (Jan. 2007), 82–96.

[65] Angelina Wang, Erin Beeghly, Sanmi Koyejo, and Daniel E Ho. 2021. Personalization in Practice: Mismatches Between User Preferences and Chatbot Behavior Reveal the Privacy Paradox and Discriminatory Double Binds.

[66] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology 54, 6 (1988), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063 Place: US Publisher: American Psychological Association.

[67] Kimi Wenzel, Nitya Devireddy, Cam Davison, and Geoff Kaufman. 2023. 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. 1–14.

[68] Marcia Farr Whiteman. 2013. Dialect influence in writing. In Writing. Routledge, 153–166.

[69] Donald Winford. 2015. The origins of African American Vernacular English. In The Oxford handbook of African American language. Oxford University Press, 85.

[70] Gloria Wong, Annie O Derthick, E J R David, Anne Saw, and Sumie Okazaki. 2014. The what, the why, and the how: A review of racial microaggressions research in psychology. Race Soc. Probl. 6, 2 (June 2014), 181–200.

[71] 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. 2025. Personalization of Large Language Models: A Survey. https://doi.org/10.48550/arXiv.2411.00027 arXiv:2411.00027 [cs].

[72] Kaitlyn Zhou, Jena D Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2024. Rel-AI: An Interaction-Centered Approach To Measuring Human-LM Reliance. arXiv preprint arXiv:2407.07950 (2024).

[73] Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. arXiv preprint arXiv:2204.03031 (2022).

A STATISTICAL TESTS

Table 2. Self-perception paired t-test results across survey settings

survey	perception	n	mean_pre	mean_post	se_pre	se_post	t	p_holm	d
text sae	positive_affect	179	4.261	4.218	0.124	0.144	0.545	0.603	-0.024
text sae	negative_affect	179	1.365	1.313	0.057	0.059	1.443	0.603	-0.067
text sae	self_consciousness	179	4.228	4.141	0.117	0.120	1.133	0.603	-0.054
text sae	self_esteem	179	5.221	5.279	0.102	0.099	-1.323	0.603	0.044
text bae	positive_affect	179	4.263	4.085	0.139	0.161	2.025	0.133	-0.088
text bae	negative_affect	179	1.494	1.423	0.068	0.064	1.097	0.548	-0.081
text bae	self_consciousness	179	4.295	4.307	0.113	0.121	-0.239	0.811	0.008
text bae	self_esteem	179	5.221	5.366	0.095	0.092	-2.970	**0.014**	0.116
speech sae	positive_affect	141	3.917	3.910	0.132	0.152	0.062	1.000	-0.004
speech sae	negative_affect	141	1.473	1.328	0.077	0.059	1.917	0.229	-0.178
speech sae	self_consciousness	141	3.965	3.941	0.121	0.137	0.441	1.000	-0.015
speech sae	self_esteem	141	5.011	5.082	0.110	0.114	-0.980	0.986	0.053
speech bae	positive_affect	120	4.094	3.956	0.148	0.181	1.099	0.383	-0.076
speech bae	negative_affect	120	1.525	1.640	0.091	0.087	-1.465	0.383	0.117
speech bae	self_consciousness	120	4.281	4.408	0.145	0.161	-1.675	0.383	0.076
speech bae	self_esteem	120	5.033	5.175	0.125	0.125	-1.680	0.383	0.103

Table 3. Model-perception Mann-Whitney U test results across survey settings

survey	perception	mean_sae	mean_bae	n_sae	n_bae	rbc	cles	p_holm	U
text	warmth	5.441	5.263	179	179	0.081	0.54	0.722	17315.5
text	competence	5.564	5.196	179	179	0.151	0.576	0.092	18446
text	trustworthy	5.123	4.872	179	179	0.098	0.549	0.704	17583
text	reliability	5.246	4.944	179	179	0.13	0.565	0.222	18102.5
text	sociability	5.218	5.101	179	179	0.025	0.513	1	16422.5
text	difficult_wordchoice	1.777	1.961	179	179	-0.084	0.458	0.722	14676.5
text	would_say	4.542	4.553	179	179	-0.024	0.488	1	15640
text	would_codeswitch	2.296	2.117	179	179	0.07	0.535	0.722	17148.5
text	difficult_phrasing	1.732	2.006	179	179	-0.086	0.457	0.722	14637.5
speech	warmth	5.56	5.108	141	120	0.218	0.609	**0.012**	10304
speech	competence	5.525	4.983	141	120	0.226	0.613	**0.01**	10374.5
speech	trustworthy	5.078	4.908	141	120	0.043	0.521	1	8822.5
speech	reliability	5.085	4.958	141	120	0.031	0.516	1	8724
speech	sociability	5.369	5.125	141	120	0.06	0.53	1	8967
speech	difficult_wordchoice	1.851	2.217	141	120	-0.078	0.461	1	7803
speech	would_say	4.128	4.433	141	120	-0.102	0.449	1	7597
speech	would_codeswitch	2.199	2.208	141	120	-0	0.5	1	8458.5
speech	difficult_phrasing	1.823	2.167	141	120	-0.076	0.462	1	7819

Table 4. BAE usage Mann-Whitney U test results across survey settings

survey	perception	mean_sae	mean_bae	n_sae	n_bae	rbc	cles	p_holm	U
text	perceived_agent_bae_usage	1.844	3.821	179	179	-0.794	0.103	**0**	3303
text	self_reported_user_bae_usage	3.084	3.48	179	179	-0.204	0.398	**0.001**	12750.5
speech	perceived_agent_bae_usage	1.73	4.192	141	120	-0.864	0.068	**0**	1153
speech	self_reported_user_bae_usage	3.234	3.517	141	120	-0.146	0.427	**0.033**	7221.5

Table 5. Dialect density t-test results for speech survey

survey	bae density	mean_sae	mean_bae	n_sae	n_bae	se_sae	se_bae	t	p_holm	d
speech	agent	0.284	2.225	141	120	0.042	0.049	-29.951	**0**	3.746
speech	user	2.092	2.35	141	120	0.071	0.078	-2.439	**0.015**	0.303

B BAE PROFICIENCY SCREENING

Now, we’d like to understand your experience with African American Vernacular English (AAVE). Please take a look at these AAVE phrases. Don’t worry about the spelling - we’re more interested in whether these phrases resonate with you. **How likely are you to say or write something in the style of the phrases below?**

Table 6. BAE Screener Phrases

Phrase	Very Likely	Likely	Somewhat Likely	Not Very Likely	Unlikely
“I know it’s a lot of ’em, I just can’t think of one”	○	○	○	○	○
“I ain’t say nothing bad though”	○	○	○	○	○
"She be tryna make me mad on purpose"	○	○	○	○	○

Received 11 September 2025