

My LLM might Mimic AAE - But When Should it?

Sandra C. Sandoval*
University of Maryland
sandracs@umd.edu

Christabel Acquaye*
University of Maryland
cacquaye@umd.edu

Kwesi Cobbina*
University of Maryland
kcobbina@umd.edu

Mohammad Nayeem Teli
University of Maryland
nayeem@umd.edu

Hal Daumé III
University of Maryland
hal3@umd.edu

Abstract

We examine the representation of African American English (AAE) in large language models (LLMs), exploring (a) the perceptions Black Americans have of how effective these technologies are at producing authentic AAE, and (b) in what contexts Black Americans find this desirable. Through both a survey of Black Americans ($n = 104$) and annotation of LLM-produced AAE by Black Americans ($n = 228$), we find that Black Americans favor choice and autonomy in determining when AAE is appropriate in LLM output. They tend to prefer that LLMs default to communicating in Mainstream U.S. English in formal settings, with greater interest in AAE production in less formal settings. When LLMs were appropriately prompted and provided in context examples, our participants found their outputs to have a level of AAE authenticity on par with transcripts of Black American speech. Select code and data for our project can be found here: <https://github.com/smelliecat/AAEMime.git>

1 Introduction

In our study, we explore how an underserved population, Black Americans in the United States, regards increasingly ubiquitous text-based AI tools in terms of their preferred functionalities and with respect to the authenticity of the language produced by these systems, given their unique needs. We specifically investigate the research questions:

1. *Do Black Americans¹ want generative AI technologies to produce African American English? If so, in what contexts?;* and
2. *How effective are large language models (LLMs) at generating authentic African American English (AAE)² when prompted to do so?*

Asterisks () indicate similar levels of contribution.

¹We use *Black Americans* to describe those who identify as American with ancestral roots to Black African ethnic groups.

²African American English is “the grammatically patterned variety of English used by many, but not all and not

The Black American population makes up approximately 13.6% of the United States total population in 2022 (U.S. Census Bureau, 2022; Moslimani et al., 2023) and represents a major stakeholder population of text-based AI technologies. AAE, while predominantly a spoken language variety, is seeing increased representation in speech-like media such as texting and social media (Blodgett et al., 2016). Its use has become synonymous with the cultural identity of some Black Americans (Bashir-Ali, 2006) with the language evolving over an extended period of time dating as far back as the period of Black enslavement in the United States. In spite of the cultural importance of AAE, Black Americans have had good reason to be hesitant to use the language outside of personal contexts due to widespread linguistic discrimination: racial identification and discrimination based on speech or writing in the work place and in schools (Baugh, 2005). In addition, the use of AAE may be associated with poverty or lower socioeconomic class (Rickford et al., 2015), which could influence Black Americans to be cautious about the circumstances under which it should be used.

AAE is predominantly identified by grammatical patterns such as the use of double negatives, variable subject-verb agreement, and omission of verbal copulas. These patterns distinguish it sharply from Mainstream U.S. English (MUSE)³. Despite its distinct linguistic characteristics and the large proportion (80-90%) of the Black American society

exclusively, African Americans in the United States” (Grieser, 2022); AAE has many alternative names, including African American Language, African American Vernacular of English, Black English, (Black) Slang, and Ebonics (Green, 2002; Wolfram and Schilling, 2015; Rickford and King, 2016; King, 2020; Becker, 2013). In our study, participants could choose the terminology they preferred.

³We refer to the most prevalent variety of American English as Mainstream U.S. English (Baker-Bell, 2020; Harris et al., 2022); as with AAE, other names exist, such as Standard American English and White Mainstream English (Wolfram and Schilling, 2015).

that speaks AAE in the United States (Holt, 2018; O’Quin, 2021; Farrington et al., 2021), prior studies have shown that AI technologies often fail to accommodate its nuances, especially in the context of speech recognition (§2).

In this paper, we investigate the research questions listed above in the context of large language models. We conduct an online study (§3) among Black Americans ($n = 104$) consisting of both a survey—to understand what is wanted out of this technology—and data annotation—to understand the efficacy of current LLMs at producing authentic AAE. In the survey (§3.1), we aim to understand in what social contexts Black Americans would want (the option of) having an LLM use AAE, including both professional and personal settings, and including both continuation behavior (e.g., email auto-complete) and reply behavior (e.g., AI assistants). In the data annotation task (§3.2), we ask Black Americans ($n = 228$ who provided 8,654 judgments for 1,357 examples) to judge text generated by three different LLMs—GPT 4o-mini, Llama 3, and Mixtral—along axes like coherence, the explicit presence of AAE features, and offensiveness. Our main contributions (§4) include:

1. We find that Black Americans favor the use of MUSE in more formal or task-specific interactions, but are open to LLM generation of AAE in personal or casual settings, preferring the autonomy to switch to AAE as desired.
2. We find that Black Americans judged the LLM AAE generations as equally authentic to the human baseline (Black American transcribed interviews), and they did not consider them to be mocking or offensive.
3. We contribute a dataset of linguistic judgments from Black American annotators on both AAE and MUSE texts, drawn both from human- and LLM-produced text. In addition, we share the dataset and a selection of our code for the project here: <https://github.com/smelliecat/AAEMime.git>

2 Related Work

Attitudes and Perceptions of AAE Speakers Towards Technology

Cunningham et al. (2024) examine the invisible labor AAE speakers undertake to be understood by language technologies. Their findings show that AAE speakers often have to proactively adapt their speech, which leads to significant frustration

and alienation when interacting with these systems. Harrington et al. (2022) further discuss how Black older adults anticipate and experience substantial challenges with voice assistants, exacerbating their reluctance to use these devices for tasks like health information seeking. Our study explores the idea that Black Americans may prefer AAE options in their interactions with speech or language-based AI technologies. We build on these lines of research by exploring Black Americans’ preferences for AAE representation in specific contextual settings.

Current State of AAE in Technology

Hill (1998) and Smokoski (2016) highlight the issue of AAE in social media as not always true AAE, but rather non-Black Americans mimicking or mocking AAE. Thus, one of our objectives for this study was to begin to understand some of the constraints that generative AAE must adhere to, to stay within the bounds of acceptability to the AAE community, such that the language is not seen as mocking individuals who use the language and remains relatable and respectful to their specific use context. Consequently, we address the idea of whether the AAE generated by LLMs could be construed as mocking or offensive in our study.

With respect to the importance of language diversity being reflected in chat-based AI technologies, Harrington et al. (2019) critique how participatory design often overlooks historical inequities, further marginalizing AAE speakers. This reflects broader systemic issues in technology design that fail to accommodate linguistic diversity.

Santiago et al. (2022) highlight the critical role of morphosyntactic features in AAE, such as the invariant ‘be’, demonstrating how leveraging these features improves the disambiguation of syntactic constructs, and mitigates the risks of discrimination in Automatic Speech Recognition (ASR) systems when misrepresented. In the context of large language models, Latimer (2023)’s ‘Black GPT’ is an example of an AI technology that has been developed with recognition of the value of AAE, and thereby promotes inclusion and representation of Black Americans (Previlon et al., 2024).

Despite recent efforts to provide inclusive technologies that meet the needs of Black Americans, Pinhanez et al. (2024) highlight the ongoing challenges to developing systems that authentically represent AAE. In particular, they look at text-to-speech systems, uncovering latent biases that prevent broad recognition and acceptance with-

out resorting to stereotypes. Issues in ASR have been documented for systems from Apple, Amazon, Google, and IBM, which exhibit error rates for Black speakers of 35%, nearly double those for White speakers at 19% (Koenecke et al., 2020).

This discrepancy highlights a systemic oversight in AI design that fails to consider AAE’s unique linguistic features, necessitating that AAE speakers frequently engage in code-switching. This adaptation requires significant invisible labor and can lead to alienation and frustration, as illustrated by Harrington’s examination of Black older adults using voice assistants (Harrington et al., 2022) and Cunningham’s insights into the emotional toll of such adjustments (Cunningham et al., 2024).

Unfortunately, the impacts of design oversights extend to chatbots and LLMs as well. Hofmann et al. (2024) expose how LLMs covertly perpetuate dialect prejudice, with their Matched Guise Probing approach, highlighting that even when trained with human feedback, these systems still enforce negative stereotypes, impacting judgments on employability and criminality.

In our study, we strive to uncover if LLMs are able to generate AAE effectively as judged by Black Americans themselves, and if they suffer from some of the same biases as speech recognition AI technologies according to our study participants. We note the distinction of the large language model-based chatbots we study here as being tasked to generate AAE, in contrast to speech recognition technologies which should understand AAE.

3 Methodology

We investigate two primary research questions: 1) whether Black Americans want Generative AI technologies to communicate and understand African American English and in which contexts, and 2) how effective LLMs are at generating AAE when prompted to do so. Our Institutional Review Board (IRB)-approved⁴ online study gathers participant feedback on these questions, via a survey on participants’ desires (§3.1), and data annotation to understand the effectiveness of current LLMs (§3.2).

For both, we recruit participants through Prolific (prolific.co), who were required to be Black American over 18 years old, familiar with AAE, and reside in the U.S. In total, of our $n = 104$

⁴Our IRB restricts us from reporting on our survey participant and annotator demographics other than at the aggregate level. Most other IRB details redacted for anonymity.

survey participants, 61 identified as female, 37 as male, 3 as non-binary, 1 as other (with 2 undisclosed). Participants ranged from 25 to 64 years old ($\mu = 34$), with the majority (81 participants) having attended at least some college. The participants come from all four major geographic areas in the U.S. (53 from the South, 21 Northeast, 17 West, and 17 Midwest) (see §A.8). For the *data annotation* tasks, we recruited $n = 228$ Black American annotators with similar demographics.

3.1 Survey of If and When Black Americans Want LLMs to Produce AAE

Our survey is designed to explore the perceptions and attitudes of Black Americans regarding AAE representation in chat-based AI systems across a variety of settings, ranging from professional to personal. For each setting, we gauge how and when participants want an LLM (or chatbot) to use AAE versus MUSE. We aimed to provide sufficient detail on the settings to make them more relatable and easier to comprehend (Lenzner, 2012). The settings are selected to give a more complete picture of Black American’s every-day experiences and preferences (Maedche et al., 2019). For each scenario presented, study participants were asked to choose from the answer choices seen in Table 1 for how they would want such an LLM to interact:

1. *AI Assistants* (professional and personal). These LLM-response settings reflect the use of an AI assistant for helping with either professional or personal tasks, and whether such an assistant should address the user in AAE.
2. *Customer Bot*. This LLM-response setting reflects the use of a text-based chatbot agent for quick assistance, and whether the agent should continue the interaction after greeting the user in AAE.
3. *Email and SMS Autocomplete*. These LLM-continuation settings reflect the use of an LLM to autocomplete a user’s own writing for emails or text messages.
4. *Educational Avatar*. This LLM-response setting reflects the use of AAE by an avatar in an education platform and whether this could impact learning experience.

3.2 Annotation of LLM Output for AAE-ness

We used text annotation to explore our second research question: how well do current LLMs generate AAE-like text (including relative to generation of MUSE-like text). We obtained human judg-

| | |
|--------------|---|
| AlwaysMUSE | LLM should always use MUSE. |
| AlwaysAAE | LLM should always use AAE. |
| AutoDetect | LLM should automatically detect/adapt to the user’s language variety. |
| UserOption | LLM should provide an option to switch between AAE and MUSE. |
| NoPreference | No preference as long as the system is effective. |

Table 1: The set of possible choices for the preferences survey, which asked Black Americans about the contexts or scenarios in which they would prefer to have language model-based AI technologies generate AAE vs MUSE.

| Description | Linguistic Judgment |
|-----------------------|--|
| <i>Coherent</i> | The text is a coherent continuation. |
| <i>AAE Features</i> | The text contains AAE features. |
| <i>Black Sounding</i> | The text sounds like something a Black American would say. |
| <i>White Sounding</i> | The text sounds like something a White American would say. |
| <i>Mocking</i> | The text is like someone mocking AAE. |
| <i>Offensive</i> | I would be offended if a chatbot said this. |

Table 2: Assessments made via Likert score rating by Black American annotators regarding the AAE and MUSE text continuations they reviewed. The continuations were either human or LLM-produced, but annotators were not told which.

ments by our Black American annotators as to how they perceived LLM-produced AAE text relative to their expectations for AAE. We show participants text, including a prefix which was transcribed (human) speech from interviews or from X (Twitter) posts, paired with a suffix which was either the natural (human) continuation or an LLM-generated continuation (by GPT, Llama, or Mixtral; see Figure 2 for examples). In some cases the human (and LLM-generated) text is AAE; in others it is MUSE. We adopt this *continuation* methodology to ensure that that LLM generations are comparable—similar topic, etc.—to the human generations to facilitate more controlled comparisons. In the annotation task, the suffix was highlighted and participants were asked to assess it along the six dimensions from Table 2 on a five-point Likert scale from “Strongly Disagree” to “Strongly Agree” (the full interface annotators saw is shown in Figure 4 in the Appendix, and see Figure 2 for examples with annotations). Participants were unaware if the suffix was human or machine-generated.

The AAE and MUSE prefixes and human baseline texts for this annotation task were drawn from the Corpus of Regional African American Language (CORAAL) (Kendall and Farrington, 2023), which is a corpus of transcribed interviews of Black Americans, an X posts (Twitter) AAE corpus (Henry, 2021), and a National Public Radio in-

terview corpus containing MUSE text (Majumder et al., 2020a,b). Hereafter, we refer to X posts as Tweets for brevity. We lightly cleaned the corpora for annotator assessments (see §A.1).

3.3 Prompting LLMs

To produce the LLM suffixes, for each prompt, we provided the following inputs to the LLMs to encourage the systems to generate completions of the interviewee statements in AAE for the CORAAL interview and X post (Twitter) prefixes, or in MUSE for the NPR interview-sourced prefixes: 1) an instruction to generate a completion (the suffix) *in AAE or in MUSE* with a suggested response length limit; 2) a list of 2-5 randomly sampled (and randomly varying in number of) training examples of authentic AAE (or MUSE for those continuations) for in context learning (Brown et al., 2020) – the AAE training examples were drawn from other CORAAL interview exchanges or other AAE Tweets, including both the interviewer statements and the corresponding interviewee responses, or the original Tweets; and 3) a test prefix as described above, that the LLM then was tasked to continue in AAE (or MUSE for the texts originally from NPR interviews). See §A.5.3 for details and §A.4 for examples of the prompt templates used.

4 Results

We outline the results from our study below. We review our findings from our scenarios-based survey, which explored Black American perspectives and expectations for interactions with language-based generative AI tools, looking at a range of personal and professional settings. We then review the results from the data annotation effort, whereby we gathered Black Americans’ judgments of the effectiveness of large language models at production of AAE and MUSE relative to our human baselines of Black American transcribed interviews – CORAAL, Tweets, and NPR Interviews.

4.1 Survey (Scenarios-based Questions)

As discussed previously in §3.1, our scenario-based questions were designed to elicit the degree to which participants wanted LLMs to use AAE in various interaction contexts. The findings from our survey reveal a nuanced interplay between user preferences and the linguistic contexts in which AI technologies might be used. Number of study participants by preference (horizontal axis) and scenario (vertical axis) as seen in Figure 1 reveal

| | | | | | | |
|----------|---------------------------|-----------|------------|------------|------------|--------------|
| Scenario | Household AI Assistant | 20 | 27 | 32 | 15 | 10 |
| | Personal AI Assistant | 13 | 30 | 35 | 21 | 5 |
| | Professional AI Assistant | 7 | 35 | 39 | 14 | 9 |
| | Educational Avatar | 8 | 40 | 36 | 12 | 8 |
| | Customer Service Chatbot | 12 | 43 | 22 | 16 | 11 |
| | SMS Autocomplete | 7 | 22 | 34 | 34 | 7 |
| | Email Autocomplete | 5 | 72 | 21 | 3 | 3 |
| | | AlwaysAAE | AlwaysMUSE | UserOption | AutoDetect | NoPreference |

Figure 1: This heatmap depicts participant ($n = 104$) preferences (horizontal axis) for the use of language varieties in seven scenarios (vertical axis). A greater number of participants preferred either for the system to use MUSE or to allow them to select between MUSE and AAE. There were some exceptions: e.g., auto-detection was considered more acceptable in SMS, and MUSE was preferred for email.

a preference gradient that spans from a strong inclination towards using MUSE in more formal or task-specific interactions to a marked openness for dialectal variability in more personalized or casual settings. This preference spectrum not only reflects current user expectations but also aligns with broader societal shifts towards more personalized and context-aware technologies.

The strong preference for MUSE in formal scenarios suggests that users prioritize consistency and efficiency, likely due to perceptions of professionalism in customer-facing AI applications. This finding is crucial as it highlights concerns about dialect prejudice in LLMs, where biases may influence AI decisions on character judgments and employability, affecting how customer inquiries are handled and responded to by AI systems. For example, Hofmann et al. (2024) emphasizes the need to address these biases, illustrating how they can impact equity and fairness in AI interactions.

Moreover, the demand for dialectal flexibility in casual or personal use scenarios underscores the importance of adaptable and culturally competent AI designs. Such adaptability is essential for ensuring that AI technologies cater to a diverse user base with varied linguistic backgrounds, thereby promoting inclusivity. This need aligns with findings from Mayfield et al. (2019), which discuss the broader impacts of NLP and AI on educational equity, stressing that technology must be sensitive to diversity to ensure equitable access to educational resources. The insights from our study suggest a nuanced approach to AI communication strategies, balancing standardization with personalization to

meet the complex preferences of users.

4.2 Annotation of AAE and MUSE texts across Six Linguistic Judgments

Our annotators made six assessments (via Likert score ratings) as seen in Table 2, providing their linguistic judgments regarding the human or LLM-generated text. This allows us to study how well LLMs are able to generate AAE-like text comparing across AAE and MUSE texts.

Mapping to Numeric Scores. We map Likert scores to the range -2 (strongly disagree) to $+2$ (strongly agree) and compute the overall score as the average across all samples.

Results Analysis Approach. We investigate how Black Americans viewed texts with regard to specific linguistic judgments. To this end, we conduct two-tailed t-tests to determine whether differences in mean scores were significantly different than zero⁵ to ascertain any statistical differences in mean Likert scores between two independent data samples at a time. Our samples were independent in that the prefixes that annotators labeled, and our annotators themselves, were non-overlapping.

We conduct two types of between-sample comparisons. The first (§4.2.1) and most critical to our approach involves *Human to Model (within corpus) comparisons of mean Likert scores*. This test is fundamentally a “does an LLM produce text like a human would, in AAE” (and “in MUSE”, as a point of reference). The second (§4.2.2), as assurance that the generated AAE is actually more like AAE than like MUSE, considers differences in judgments between AAE text and MUSE text (as opposed to human text vs. machine text).

4.2.1 How well do LLMs generate AAE? Analysis of Black Americans’ Ratings of LLM versus AAE suffixes

In our first analysis, we compare the mean Likert scores for human baseline texts for each corpus (i.e., the human continuation of a text) to the model

⁵Unless otherwise noted, we take “significant” to mean false discovery rate of 5%. We apply Bonferroni corrections to our p-values, reporting 95 percent confidence intervals, since we performed multiple t-tests assessing differences between Likert score means between samples for each linguistic judgment (within and between corpora for each annotation survey we administered). Bonferroni corrections conservatively report statistical significance. We had six within-corpus comparisons and four between-corpus comparisons per label per each of the two annotation surveys, resulting in a total of 10 comparisons used for each judgment’s bonferroni correction.

| System | Coherence | | | AAE Features | | |
|---------|-----------------|-----------------|-----------------|--------------------------------|------------------|--------------------------------|
| | CORAAL | NPR | TWEETS | CORAAL | NPR | TWEETS |
| Human | 0.35 \pm 0.30 | 0.99 \pm 0.51 | 0.04 \pm 0.54 | 0.18 \pm 0.39 | -0.55 \pm 0.84 | -0.57 \pm 0.87 |
| GPT | 0.80 \pm 0.07 | 1.29 \pm 0.36 | 0.64 \pm 0.24 | 1.18 ^{***} \pm 0.17 | -0.61 \pm 0.87 | 0.99 ^{***} \pm 0.07 |
| Llama | 0.66 \pm 0.20 | 1.31 \pm 0.34 | 0.43 \pm 0.35 | 0.86 [*] \pm 0.03 | -0.67 \pm 0.90 | 0.57 ^{***} \pm 0.29 |
| Mixtral | 0.60 \pm 0.15 | 1.21 \pm 0.44 | 0.48 \pm 0.33 | 0.72 \pm 0.11 | -0.67 \pm 0.90 | 0.60 ^{***} \pm 0.24 |

| System | Black Sounding | | | White Sounding | | |
|---------|-------------------------------|-----------------|--------------------------------|-------------------------------|-----------------|------------------|
| | CORAAL | NPR | TWEETS | CORAAL | NPR | TWEETS |
| Human | 0.39 \pm 0.28 | 0.15 \pm 0.42 | -0.30 \pm 0.72 | -0.23 \pm 0.62 | 0.83 \pm 0.42 | -0.65 \pm 1.00 |
| GPT | 1.01 ^{**} \pm 0.07 | 0.21 \pm 0.38 | 0.79 ^{***} \pm 0.13 | -0.83 [*] \pm 0.88 | 0.89 \pm 0.39 | -1.07 \pm 1.21 |
| Llama | 0.85 \pm 0.02 | 0.31 \pm 0.34 | 0.22 \pm 0.45 | -0.74 \pm 0.87 | 1.02 \pm 0.32 | -0.96 \pm 1.16 |
| Mixtral | 0.85 \pm 0.04 | 0.11 \pm 0.43 | 0.37 [*] \pm 0.37 | -0.49 \pm 0.75 | 0.81 \pm 0.43 | -1.06 \pm 1.21 |

| System | Mocking | | | Offensive | | |
|---------|------------------|------------------|--------------------------------|------------------|------------------|---------------------------------|
| | CORAAL | NPR | TWEETS | CORAAL | NPR | TWEETS |
| Human | -0.88 \pm 1.35 | -1.47 \pm 2.25 | -0.79 \pm 1.21 | -0.78 \pm 1.19 | -1.48 \pm 2.27 | -0.96 \pm 1.47 |
| GPT | -0.57 \pm 1.19 | -1.61 \pm 2.32 | -0.57 \pm 1.27 | -0.86 \pm 1.23 | -1.60 \pm 2.33 | -0.57 \pm 1.27 |
| Llama | -0.46 \pm 1.13 | -1.63 \pm 2.33 | 0.14 ^{***} \pm 0.74 | -0.53 \pm 1.07 | -1.56 \pm 2.25 | -0.07 ^{***} \pm 1.02 |
| Mixtral | -0.74 \pm 1.28 | -1.60 \pm 2.31 | -0.20 \pm 0.91 | -0.86 \pm 1.24 | -1.60 \pm 2.33 | -0.55 \pm 1.26 |

Table 3: Mean Likert scores for each LLM for a given linguistic judgment and corpus (n ranged from 119 to 126 Likert score observations for a given sample in a two sample comparison; 72 two-sample comparisons were conducted.). Scores for the original human text are shown in the Human row. $p < 0.05$ is marked with *; $p < 0.01$ with ** and $p < 0.001$ with ***.

continuations for prefixes from that corpus as seen in Table 3. For these tests, mean Likert scores are assumed to be equal (have zero difference in means between the two samples in question).

In general, study participants rated the highlighted suffixes from the AAE produced by all three LLMs, GPT, Llama and Mixtral, equally or better than the original AAE texts (human baselines) across the linguistic judgment assessments of: 1) text continuation coherency, 2) the texts containing AAE features, and 3) the texts sounding like something a Black American might say. Specifically, their Likert scores were in the Agreement range ($\mu > 0$) for these three judgments for the AAE generations they annotated, which is positive for Black Americans favoring AAE or language choice in LLMs (the AlwaysAAE, UserOption and AutoDetect columns of Figure 1).

For the first linguistic judgment *Coherence*, Black Americans agreed that the original human CORAAL AAE text as well as the model continuations were coherent (the difference in means was not statistically significant), with Tweets considered slightly less coherent hovering around neutral (μ close to 0) for the original posts and slightly agree for the model generations. Regarding the second linguistic judgment *AAE Features*, while annotators considered some model-generated texts for CORAAL ($\mu = 1.18$ for GPT and $\mu = 0.86$ for Llama) to possess a greater extent of AAE features

in comparison to the CORAAL AAE human baseline ($\mu = 0.18$), they viewed both as displaying features indicative of AAE. In contrast, annotators *disagreed* that the original Tweets ($\mu = -0.57$) contained AAE features unlike the model-generated-Tweets, where all models were rated more highly, particularly GPT ($\mu = 0.99$).

Similar to the judgment on AAE feature presence, annotators generally agreed for the third linguistic judgment, abbreviated *Black Sounding*, that the CORAAL suffixes read like something a Black American would say, with slight agreement for the human texts at $\mu = 0.39$ and stronger agreement for the model-generated text across systems, where GPT had the highest mean $\mu = 1.01$ that was also statistically different from the human baseline text. Again, the Twitter human texts ($\mu = -0.30$) were judged to *not* sound like something a Black American would say whereas annotators agreed that the model-generated Twitter-continuations resembled something a Black American might say. $\mu = 0.37$ for Mixtral and $\mu = 0.79$ for GPT showed a statistical difference from the human baseline Tweets.

The last three linguistic judgments annotators were asked to assess would be less favorable if Black Americans agreed with them for the AAE texts; these included: 1) the text sounded like something a White American would say, 2) the text could be perceived as mocking how some Black Americans speak, and finally 3) the text could be con-

strued as offensive coming from a chatbot. Likert ratings from these judgments of the AAE texts were largely on the “disagree” side ($\mu < 0$).

For our fourth linguistic judgment – *White Sounding*, Annotators felt strongly that the highlighted AAE suffixes *did not* sound like something a White American would say. All scores, whether for the AAE human baselines ($\mu = -0.23$ for CORAAL and $\mu = -0.65$ for Tweets) or for the model-generated AAE continuations, were in the disagreement range, with GPT showing a statistically different mean of $\mu = -0.83$ for CORAAL generations relative to the CORAAL human text.

For the fifth linguistic judgment *Mocking*, annotators disagreed for both human (with $\mu = -0.88$ for CORAAL and $\mu = -0.79$ for Tweets) and most model-generated AAE (ranging from $\mu = -0.74$ for CORAAL Mixtral generations to $\mu = 0.13$ for Llama Tweets generations), that the texts sounded like someone making fun of the way some Black Americans speak. The only statistical difference in means relative to the Tweets-human baseline was for Llama generations; for these, annotators slightly agreed that the model generations could be perceived as making fun of the way some Black Americans speak. Annotators more strongly disagreed with this judgment regarding the MUSE human and model-generated texts, perceiving both as not like mocking Black Americans and assessing them equivalently in their Likert scores.

For our final linguistic judgment, annotators generally disagreed that they would be offended by either the CORAAL or Tweets human or model-generated texts with an exception for Tweets, where they felt neutral that the model-generated text for Llama would be offensive ($\mu = -0.07$) relative to the human posts which they disagreed ($\mu = -0.96$) would be offensive. They yet more strongly disagreed that the MUSE human or model-generated texts could be interpreted as offensive, judging them roughly equivalent in Likert scores.

For all six judgment types for the human to model comparisons, most differences in means between a given corpus human baseline and an LLM generation for that corpus were not statistically different from zero except those mentioned for the AAE corpora. Model-generated MUSE was not statistically different from the human MUSE.

To clarify, the NPR corpus was included as a proxy for MUSE, the predominant version of English for communicating in the United States, to facilitate understanding judgments of the AAE texts

relative to MUSE. We would expect LLMs to be proficient in the production of MUSE, and it is understood that LLMs in the United States default to MUSE. While the human-to-model comparisons within each AAE corpus could provide insight into how well an LLM generates AAE, given examples of it, we also must assess ratings of AAE texts relative to the NPR MUSE baseline.

In general, as one might expect, our study’s Black American annotators agreed most that the MUSE human baselines and model generations had the most coherent continuations, were most White American sounding, and disagreed that the MUSE texts contained AAE features and sounded like something a Black American would say. Conversely, the AAE continuations were considered to be most like something a Black American would say and as having AAE features, but it is worth noting that Black Americans also felt that the MUSE continuations could sound like something a Black American might say. This suggests that while AAE is distinct and should be treated as such, Black Americans speak a wide range of English and these findings highlight Black Americans’ bidialectal nature, also previously highlighted in §2 where we discuss Black Americans’ needs to code-switch and dialect prejudice in LLMs.

4.2.2 How Black Americans Perceived AAE versus MUSE texts

The previous results showed, roughly, that the LLM-generated AAE was on par with human-written AAE across many linguistic axes. Thus, we seek to ensure that the models are actually generating AAE (versus MUSE) when prompted to do so, that the human text was actually AAE (versus MUSE), and that people could tell the difference. To answer this question, we conduct between-corpus tests of *AAE to MUSE* as seen in Table 4 (our previous tests only compared human to model-generated texts within one corpus at a time).

Similar to the previous analysis, we compared two sample means at a time with t-tests, one from an AAE corpus and another from the MUSE corpus (e.g., CORAAL AAE human baseline compared to MUSE human baseline, or Twitter GPT continuation compared to MUSE GPT continuation). In contrast, in testing between two dialects, MUSE and AAE, we expected the alternative hypothesis – a non-zero difference in mean Likert scores between the two samples – to be true.

Table 4 results show that across our set of linguis-

| System | Coherent | | AAE Feats | | Black-Snd | | White-Snd | | Mocking | | Offensive | |
|---------|----------|-----|-----------|-----|-----------|----|-----------|-----|---------|-----|-----------|-----|
| | Co | Tw | Co | Tw | Co | Tw | Co | Tw | Co | Tw | Co | Tw |
| Human | ** | ** | ** | - | - | - | *** | *** | * | ** | ** | - |
| GPT | - | *** | *** | *** | *** | * | *** | *** | *** | *** | *** | *** |
| Llama | ** | *** | *** | *** | * | - | *** | *** | *** | *** | *** | *** |
| Mixtral | - | *** | *** | *** | ** | - | *** | *** | *** | *** | *** | *** |

Table 4: Statistical significance indicated for 48 between-corpus comparisons of mean Likert scores (n ranged from 119 to 126 Likert score observations for a given sample in a two sample comparison), where values shown resulted from each t-test comparing an AAE corpus versus the MUSE corpus of NPR Interviews, latter not labeled. We indicate Linguistic judgment and the AAE Corpus (CORAAL or Tweets) for each t-test. $p < 0.05$ is marked with *; $p < 0.01$ with ** and $p < 0.001$ with ***.

tic judgments (as seen in Table 2), when we compare MUSE against CORAAL or against Tweets, virtually all between MUSE and AAE comparisons show a statistically significant difference in means; in other words, Black Americans perceived the AAE human-originated texts and the model-generated texts as distinct from the MUSE texts. The statistical significance of the comparisons of MUSE versus CORAAL, or MUSE versus Tweets, Likert scores for each linguistic judgment can be seen in the columns, alternating within each judgment between CORAAL versus Tweets.

4.3 Discussion of Linguistic Judgments

For transparency, we initially had relatively low expectations for the capabilities and performance of the text-based generative AI models with respect to AAE, given the issues we outlined regarding AI technologies’ abilities to understand or process spoken language (§2), as well as the likelihood that AAE has a distinct minority representation in the training corpora for these systems. Therefore, one of our most noteworthy results is contrary to our expectations: that LLMs in general performed similarly to our human baseline, and in some cases were actually seen as containing more AAE features or sounding more like something a Black American might say than our human baseline AAE texts. The suffixes for all three LLMs were judged being more coherent or easier to understand and flow better from the prefix than the human baseline. Furthermore, it is encouraging that the suffixes generated by LLMs were on average judged to be inoffensive, not mocking of Black Americans, and not White Sounding. However, “on average” does not mean that none of the outputs were problematic.

As seen in Figure 2, there may be a minority of cases where generative AI may produce text containing AAE that is undesirable for any number of reasons (whether because of the nature of the AAE in it, the text content or other factors), and different people may respond negatively or pos-

itively to the same text given inherently varying perspectives. In the worst cases, AAE generated could perpetuate stereotypes, mock Black Americans, or otherwise generate inauthentic AAE. However, we have shown that LLM systems generally do not seem to be doing that, and we believe that there are meaningful and highly impactful benefits to be gained from the generation of AAE in popular language-based technologies, such as the increased representation of Black American expression through AAE and the promotion of inclusivity and improved quality of service for them as stakeholders. In our opinion, these gains would outweigh the risks, and extrapolation of our results (see Table 3 and Figure 1) would indicate that the large population of Black Americans (§1) would support this, preferring AAE in more informal contexts but wanting the autonomy to choose between AAE and Mainstream U.S. English (MUSE). In fact, our approach may be generalized to hundreds of dialects of English to verify and promote their acceptance in LLM products. Please see our motivation for dialectical diversity in §2. If Black Americans engage more directly in AAE, when given the choice, there would be limited risk of the AAE generations by LLMs becoming more artificial since LLMs already produce credible AAE. With a greater number and variety of Black American users to train LLM systems, production of AAE will increasingly meet Black Americans’ expectations. Of course, AAE could have negative impacts if the LLM-generated AAE is inconsistent with Black American user preferences for AAE. However, in our survey, Black Americans unequivocally expressed that there are contexts in which they are interested in AAE generations, wanting the freedom to choose this as desired. Thus, with appropriate safeguards to avoid offensive or mocking text generation, these risks could be well-mitigated.

| |
|---|
| <p>Prefix Interviewer: [Right, right. <laugh>] Interviewee: You know, a lot of people got their Archie Bunker [and are k- in Queens.], and lots of drugs. So it became</p> <p>Suffix: real hard fo' me tuh concentrate, ya feel me? But ah managed tuh keep mah grades up, even though ah wuz livin' in dat kinda environment. Ah mean, education wuz always important tuh me, so ah made su' tuh prioritize it, no matter what else wuz happenin'. Dat's why ah say - it ain' jus 'bout where yuh from, it's 'bout where yuh want tuh go in life. An' ah knew ah wanted tuh go far, ya dig?</p> <p>Response Generated by: Mixtral-8x7B Linguistic Judgment: The highlighted text sounds like something someone would say making fun of the way some Black Americans speak. Annotation: 2 - Strongest Agreement</p> |
| <p>Prefix Interviewer: Yeah. So do you like the school that, uh, /RD-NAME-2/'s at right now? Interviewee: [Mm-hm.] Oh yeah, that's what I told her, I said- like we were looking at the rental house. She's like, oh my god. I</p> <p>Suffix: I said, "Girl, this ain't nothing." I mean, she thought it was rough, but I grew up in some tough neighborhoods too, you feel me? I told her, "You gotta see the beauty in it," 'cause once you start seeing the beauty, you start appreciating the struggle. And that's what makes us strong, you know what I'm saying? We came from nothing, built ourselves up, and now we're trying to build our own empire, you hear me? So yeah, I love that school, I love what it represents, resilience, perseverance... that's the DC way</p> <p>Response Generated by: Meta-Llama-3-70B-Instruct Linguistic Judgment: The highlighted text sounds like something someone would say making fun of the way some Black Americans speak. Annotation: 2 - Strongest Agreement</p> |
| <p>Prefix Interviewer: Yeah. So do you like the school that, uh, /RD-NAME-2/'s at right now? Interviewee: [Mm-hm.] Oh yeah, that's what I told her, I said- like we were looking at the rental house. She's like, oh my god. I</p> <p>Suffix: was like, "Girl, you ain't seen nothin' yet!" I love the vibe there, real good community feel, and the teachers, they be lookin' out for the kids. I told her, "You gon' thrive here, no doubt!" Plus, the after-school programs, they got a little somethin' for everybody, music, sports, you name it. It's a place where they celebrate our culture and they encourage the kids to express themselves. That's important, you know? I just want her to soak it all up and enjoy the ride. This is a whole new chapter for us, and I'm just tryna make sure she gets the most outta it!</p> <p>Response Generated by: GPT 4o-mini Linguistic Judgment: I would be offended if a chatbot said something like the highlighted text. Annotation: 2 - Strongest Agreement</p> |

Figure 2: Examples of response continuations generated by Mixtral, Llama, and GPT, with annotation scores based on human participants' linguistic judgments.

5 Conclusion

In this study, we explore how Black Americans perceive the appropriate use of AAE pertaining to language-model-based technologies, especially in terms of their ability to authentically represent AAE. We consider both the expectations of the community and the actual output of current systems, exploring the idea that AAE should not merely be an option but a well-integrated feature.

Overall, the LLMs we reviewed were surprisingly capable and comparable in how they were perceived in terms of their ability to produce authentic AAE in comparison to the transcribed speech of Black Americans from the CORAAL corpus. We found that the text completions by some LLMs were often perceived as more AAE-heavy, or sounding more like something a Black American would say, than our the language in our human AAE cor-

pus. If our human AAE baseline is assumed to have the "right amount" of AAE, then having more linguistic features of AAE than the baseline could be considered to be excessive by some AAE speakers, whereas less may be an insufficient amount of AAE. At the same time, because the human AAE is a transcript, it may not be fully reflective of all ways that AAE is used in practice. LLMs are either slightly under-doing or over-doing AAE, but on average, participants generally disagreed that the machine-generated text by the LLMs we studied was offensive to or mocking of Black Americans.

For the scenarios-based questions of our survey, our findings reveal intricate preferences for AI applications across diverse environments, indicative of broader societal shifts toward technologies that are both personalized and context-sensitive. These insights are pivotal for developers and policymakers tasked with refining AI tech to align more closely with user expectations, thereby facilitating smoother integration of AI into everyday life.

Future Work Our study highlights that Black Americans prefer, at a minimum, the option for communication in AAE with popular language-based Generative AI tools and generally deem LLM-generated AAE as credible and similar to spoken AAE. In light of this finding, we encourage the technology community to expand the linguistic diversity in language-based Generative AI tools; in particular, they should consider functionality that provides the autonomy to Black Americans to switch to AAE on demand in circumstances of their choosing, as well as have an array of multimodal options for AAE communications, including but not limited to generation and understanding of AAE text as well as audio communications, including speech recognition and production. More generally, technological support for alternative dialects or sociolects in generative AI systems will make these systems more broadly accepted and equitable for a range of important stakeholder populations. Finally, we encourage future research on the relationship between 1) the diverse attributes that characterize Black Americans (whether they be regional, cultural, socioeconomic or demographic) and 2) whether and how they express AAE personally, or their preferences for its production by Generative AI tools across different contexts.

6 Ethics Statement

We recruited Black American study participants to provide their opinions about the generation of AAE in AI technologies, such as AI assistants, and make judgments about how effectively these systems produce AAE. We do not believe our study participants were exposed to any meaningful risks through this process, and we ensured that their remuneration was fair and above average (two and a half times the U.S. federal minimum wage) for their time. Any minor risks that our participants might have been exposed to were delineated in our application to the Institutional Review Board of *redacted*, which was approved with a status of “Ex-empt” on *redacted*. All study participants provided informed consent for their participation. All data utilized by the large language models in this study was anonymized; specifically, we used publicly available transcriptions of interviews with Black Americans from the CORAAL corpus, which was anonymous when we retrieved it online. Finally, we utilized AI code-writing assistance to develop our code used to prepare our data sets.

7 Limitations

Data Limitations. As in most data annotation tasks, we were limited by the data available to us. We chose to work with a corpus of transcribed interviews of Black Americans from the reputable CORAAL online repository to represent authentic AAE to our best ability. We chose this corpus because it was based on conversation and not written language (AAE is most commonly spoken), was informal and likely to have more AAE than most transcribed interviews due to the fact that interviews were often amongst acquaintances including friends and community members, and the corpus was rich in regional variation having both male and female interviewees represented. Even so, there were many cases of annotations in the exchanges labeled by our study participants (for punctuation or laughter, for example) which may have seemed awkward or confusing. These annotations were made by the CORAAL data stewards via the transcription and editing process, and could have influenced how authentic or coherent the text was perceived. Furthermore, there was a great range in the extent of AAE linguistic features we observed in the transcribed speech between interviews; given this, even though we randomly sampled which exchanges would be labeled by annotators, some an-

notators may have been exposed to more or less AAE in the exchanges than others.

Researcher positionality. When this manuscript was drafted, one author self-identified as a bicultural White American Latina female who does not speak AAE, one identified as an African female who does not speak AAE, one identified as Black African male who is familiar with AAE, one identified as a Kashmiri male who does not speak AAE, and one identified as a White male who does not speak AAE. Our background and positionality has limited our direct, personal understanding of Black American preferences for how generative AI technologies should perform regarding AAE. We aimed to mitigate our limitations by soliciting feedback from Black Americans in the pilot phase of the survey, and also by limiting our survey study respondents to only Black American adults.

Participant Limitations. All languages including AAE are complex, and by having annotators label single pairs of statements, certainly some of the nuance of language is lost that might otherwise be present in a full dialogue between Black American speakers. Furthermore, while our annotator pool consisted of a relatively diverse set of Black Americans, they tended toward more educated and may have been unrepresentative in other ways we did not measure (as is typical for online crowdsourced studies). Our IRB also restricted our reporting on study participant and annotator demographics other than at the aggregate level. To ensure statistically significant insights, we focused our analysis on the groups of participants as a whole.

LLM Limitations. Finally, the LLM generations that resulted from our prompting, no matter how careful, are inherently limited by the text corpora upon which they are trained. Our ability to “get the LLM to use AAE” is limited by our ability to prompt the models well; it is possible—indeed likely—that alternative prompts would lead to substantially different results.

8 Acknowledgments

We would like to thank the University of Maryland (UMD) CLIP Lab’s students and other UMD students who supported our pilot studies, the Black Americans who provided invaluable perspectives through our survey and annotation effort, and Professors Nicole Holliday and Shenika Hankerson

as well as Jay Cunningham for their helpful insights regarding African American English early in the project. This material is based upon work partially supported by the NSF under Grant No. 2131508 and Grant No. 2229885 (NSF Institute for Trustworthy AI in Law and Society, TRAILS). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

- AI@Meta. 2024. [Llama 3 model card](#).
- April Baker-Bell. 2020. Dismantling anti-Black linguistic racism in English language arts classrooms: Toward an anti-racist Black language pedagogy. *Theory Into Practice*, 59(1):8–21.
- K. Bashir-Ali. 2006. [Language learning and the definition of one’s social, cultural and racial identity](#). *TESOL Quarterly*, 40(3):628–639.
- John Baugh. 2005. Linguistic profiling. In *Black linguistics*, pages 167–180. Routledge.
- Kara Becker. 2013. Ethnolect, dialect, and linguistic repertoire in New York City. *Manuscript Submitted for Publication*. In *Malcah Yaeger-Dror and Lauren Hall-Lew (eds.), New Perspectives on the Concept of Ethnolect: Publication of the American Dialect Society (Pads) Durham: Duke University Press*.
- 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*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. [Chatbot arena: An open platform for evaluating LLMs by human preference](#). *Preprint*, arXiv:2403.04132.
- Jay L. Cunningham, Su Lin Blodgett, Hal Daumé III, Christina Harrington, Hanna Wallach, and Michael Madaio. 2024. Understanding the impacts of language technologies’ performance disparities on African American language speakers. In *Proceedings of the Conference on Language Technologies*, Location of the Conference. University of Washington; Microsoft Research; Carnegie Mellon University; Google Research.
- C. Farrington, S. King, and M. Kohn. 2021. [Sources of variation in the speech of African Americans: Perspectives from sociophonetics](#). *Wiley Interdisciplinary Reviews: Cognitive Science*, 12(3):e1550.
- Lisa J Green. 2002. *African American English: a linguistic introduction*. Cambridge University Press.
- J.A. Grieser. 2022. *The Black Side of the River: Race, Language, and Belonging in Washington, DC*. G - Reference, Information and Interdisciplinary Subjects Series. Georgetown University Press.
- Christina N. Harrington, Sheena Lewis Erete, and Anne Marie Piper. 2019. [Deconstructing community-based collaborative design](#). *Proceedings of the ACM on Human-Computer Interaction*, 3:1 – 25.
- 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 *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI ’22, New York, NY, USA. Association for Computing Machinery.
- Camille Harris, Matan Halevy, Ayanna Howard, Amy Bruckman, and Diyi Yang. 2022. Exploring the role of grammar and word choice in bias toward African American English (AAE) in hate speech classification. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 789–798.
- Jazmia Henry. 2021. [Aave corpora](#). Accessed: 2024-06-01.
- Jane H Hill. 1998. Language, race, and white public space. *American anthropologist*, 100(3):680–689.
- Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. [Dialect prejudice predicts AI decisions about people’s character, employability, and criminality](#). *Preprint*, arXiv:2403.00742.
- Yolanda Holt. 2018. [Mechanisms of Vowel Variation in African American English](#). *Journal of Speech Language and Hearing Research*, 61:1.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Guillaume Lengyel, Guillaume Bour, Guillaume Lample, L  lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th  ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2024. [Mixtral of experts](#). *Preprint*, arXiv:2401.04088.
- Tyler Kendall and Charlie Farrington. 2023. [The Corpus of Regional African American Language](#).

- Sharese King. 2020. From African American Vernacular English to African American Language: Rethinking the study of race and language in African Americans' Speech. *Annual Review of Linguistics*, 6:285–300.
- A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, et al. 2020. **Racial disparities in automated speech recognition**. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.
- Sonja L. Lanehart, Jennifer Bloomquist, and Ayesha M. Malik. 2015. **Language Use in African American Communities: An Introduction**. In *The Oxford Handbook of African American Language*. Oxford University Press.
- Latimer. 2023. **LATIMER'S BLACK GPT**.
- Timo Lenzner. 2012. Effects of survey question comprehensibility on response quality. *Field Methods*, 24(4):409–428.
- A. Maedche, Christine Legner, Alexander Benlian, Benedikt Berger, Henner Gimpel, T. Hess, O. Hinz, Stefan Morana, and Matthias Söllner. 2019. **AI-Based Digital Assistants**. *Business & Information Systems Engineering*, 61:535–544.
- Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020a. Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. *Equal Contribution.
- Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020b. **A large-scale open-source corpus of media dialog**. *CoRR*, abs/2004.03090. *Equal Contribution.
- Elijah Mayfield, Michael Madaio, Shrimai Prabhumoye, David Gerritsen, Brittany McLaughlin, Ezekiel Dixon-Román, and Alan W Black. 2019. **Equity beyond bias in language technologies for education**. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 444–460, Florence, Italy. Association for Computational Linguistics.
- M. Moslimani, C. Tamir, A. Budiman, L. Noe-Bustamante, and L. Mora. 2023. **Facts About the U.S. Black Population**.
- Camille O'Quin. 2021. *Exploring African American Vernacular English and Disproportionality in Special Education*. Phd thesis, Illinois State University.
- Claudio Santos Pinhanez, Raul Fernandez, Marcelo Carpinette Grave, Julio Nogima, and Ron Hoory. 2024. **Creating an African American-Sounding TTS: Guidelines, Technical Challenges, and Surprising Evaluations**. In *Proceedings of the 29th International Conference on Intelligent User Interfaces, IUI '24*, page 259–273, New York, NY, USA. Association for Computing Machinery.
- Wilermine Previlon, Alice Rozet, Jotsna Gowda, Bill Dyer, Kevin Tang, and Sarah Moeller. 2024. **Leveraging syntactic dependencies in disambiguation: The case of African American English**. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10403–10415, Torino, Italia. ELRA and ICCL.
- John R Rickford, Greg J Duncan, Lisa A Gennetian, Ray Yun Gou, Rebecca Greene, Lawrence F Katz, Ronald C Kessler, Jeffrey R Kling, Lisa Sanbonmatsu, Andres E Sanchez-Ordoñez, et al. 2015. Neighborhood effects on use of African-American Vernacular English. *Proceedings of the National Academy of Sciences*, 112(38):11817–11822.
- 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*, pages 948–988.
- Harrison Santiago, Joshua Martin, Sarah Moeller, and Kevin Tang. 2022. Disambiguation of morpho-syntactic features of African American English—the case of habitual be. *arXiv preprint arXiv:2204.12421*.
- Hanna L Smokoski. 2016. Voicing the other: mock aave on social media.
- U.S. Census Bureau. 2022. **Quickfacts: United states**.
- Walt Wolfram and Natalie Schilling. 2015. *American English: dialects and variation*. John Wiley & Sons.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, and Yongqiang Ma. 2024. **Llamafactory: Unified efficient fine-tuning of 100+ language models**. *arXiv preprint arXiv:2403.13372*.

A Appendix

A.1 Preparation of the CORAAL Corpus (Black American transcribed interviews)

Our prompt texts or prefixes needed to have authentic AAE to the degree possible and cover a broad range of the different variations of AAE spoken in the wild. (Lanehart et al., 2015). To achieve this goal, we made use of the CORAAL interview transcriptions, choosing 30 interviews from more than 220 transcribed interviews available (interviews were conducted with AAE speakers born between 1888 and 2005). The CORAAL AAE speakers who participated in these interviews are from six (6) cities across the United States with large Black populations, including: Washington, D.C. (from 1968 and 2016 interviews), Detroit, Michigan, Lower East Side New York City, New York, Princeville, North Carolina, Rochester New York and Valdosta, Georgia (Kendall and Farrington, 2023). The thirty (30) interviews mentioned above (Selected Interviews) from the CORAAL corpus were chosen to be balanced by sex and randomly sampled by location (but ensuring that we drew from the cities mentioned above). We considered these original interviews of Black Americans from the CORAAL corpus to be our human baseline or the AAE “ground truth”; in other words, this corpus was considered to represent authentic AAE spoken by Black Americans, to investigate our original research questions.

We proposed an initial set of criteria for a valid interview in our setting, which was that it must include only two (2) participants and any given interviewer and interviewee statements from the CORAAL corpus must have been greater than five tokens (words) long to be included (since short utterances were typically filler words such as "uh huh" or similar acknowledgments), unless they contained "who, what, where, when and why" types of questions, which contained relevant content. Additionally, any pairs of exchanges between the interviewer and the interviewee, where the interviewee’s response was less than 20 tokens (words) long, were excluded since it was important to have sufficiently long prefixes for the LLMs to create coherent and meaningful continuations of the interviewee response. To collect the human judgments on the LLM-generated texts relative to our human baseline, via the second part of our online study we provided our study participants approximately eight conversational exchanges per person, where each exchange consisted of an interviewer statement followed by the interviewee response.

A.2 Selected Interviews

ATL_textfiles_2020.05/ATL_se0_ag2_f_01_1.txt,
DCB_textfiles_2018.10.06/DCB_se1_ag1_f_01_1.txt,
DCB_textfiles_2018.10.06/DCB_se2_ag1_m_01_1.txt,
DTA_textfiles_2023.06/DTA_se1_ag3_m_02_1.txt,
LES_textfiles_2021.07/LES_se0_ag3_m_01_1.txt,
PRV_textfiles_2018.10.06/PRV_se0_ag2_m_02_1.txt,
ROC_textfiles_2020.05/ROC_se0_ag2_f_04_1.txt,
ROC_textfiles_2020.05/ROC_se0_ag2_m_01_1.txt,
ROC_textfiles_2020.05/ROC_se0_ag3_f_02_1.txt,
VLD_textfiles_2021.07/VLD_se0_ag2_f_01_1.txt,
ATL_textfiles_2020.05/ATL_se0_ag1_f_01_1.txt,
DCA_textfiles_2018.10.06/DCA_se1_ag1_f_02_1.txt,
DCB_textfiles_2018.10.06/DCB_se1_ag2_f_01_1.txt,
DTA_textfiles_2023.06/DTA_se1_ag1_f_01_1.txt,
LES_textfiles_2021.07/LES_se0_ag2_f_01_1.txt,
PRV_textfiles_2018.10.06/PRV_se0_ag1_f_01_2.txt,
ROC_textfiles_2020.05/ROC_se0_ag1_f_02_1.txt,
VLD_textfiles_2021.07/VLD_se0_ag3_f_01_2.txt,
DTA_textfiles_2023.06/DTA_se1_ag1_f_02_1.txt,
ROC_textfiles_2020.05/ROC_se0_ag1_f_03_1.txt,
ATL_textfiles_2020.05/ATL_se0_ag1_m_04_2.txt,
DCA_textfiles_2018.10.06/DCA_se1_ag3_m_01_1.txt,
DCA_textfiles_2018.10.06/DCA_se3_ag4_m_01_1.txt,
DCB_textfiles_2018.10.06/DCB_se3_ag3_m_02_1.txt,
DTA_textfiles_2023.06/DTA_se1_ag1_m_01_1.txt,
DTA_textfiles_2023.06/DTA_se2_ag4_m_02_1.txt,
LES_textfiles_2021.07/LES_se0_ag4_m_01_1.txt,
VLD_textfiles_2021.07/VLD_se0_ag2_m_01_1.txt,
VLD_textfiles_2021.07/VLD_se0_ag3_m_02_1.txt,
DCB_textfiles_2018.10.06/DCB_se1_ag2_m_02_1.txt

A.3 Prompting LLMs

The LLMs in our study were prompted to create continuations or text completions of the original interviewee statements. These, along with the responses of the interviewees from our human baseline (CORAAL), were later annotated by our study participants. The process of creating LLM continuations of interviewee statements involved first converting the CORAAL interviewer and interviewee exchanges into a format suitable for LLM input. This included systematic editing to alternate responses between the interviewer and interviewee, to maintain some flow and coherency in the conversation. We also removed non-linguistic features like "<pause>" as outlined in the CORAAL online corpus documentation, so we could focus more on the important linguistic features.

To generate our LLM outputs, we opted to use three of the most popular advanced LLMs. Namely OpenAI GPT 4o-mini, Meta-Llama-3-70B-Instruct and Mixtral-8x7B-Instruct-v0.1 (Brown et al., 2020; AI@Meta, 2024; Jiang et al., 2024). The choice of these models was based on their leading performance in natural language processing tasks and their widespread adoption (Chiang et al., 2024). We utilized a custom system prompt (Final System Prompts) for all 3 models. This system prompt included instructions on the objective of the task and guidelines on how the models were expected to respond to the user prompts. The OpenAI prompt was performed with the OpenAI API, while the open weights models (LLAMA and Mixtral) were prompted with a modified version of the Llama factory code base (Zheng et al., 2024). Subsequently, we explored 3 prompting strategies. We found that, in the zero-shot setting, all three models provided responses that failed to follow the instruction, refused to answer user prompts, or provided responses that did not fit in the context of the prompt (Zero-Shot Example). Subsequently, we experimented with providing chat history to the models (i.e. using an "in context learning" approach (Brown et al., 2020)). In this setting, we tested two different sources from the chat history. The first was from the model responses, where we kept the models' own responses and added them to the new prompts as chat history. This performed slightly better than the zero-shot setting but still struggled to stay in context of the conversation as once the model deviated, the entire conversation followed in the said deviation. This happened quite often since the first few prompts always had a chat history with few irrelevant chat histories (salutations) (Model History Example). The second and best-performing setting was using the CORAAL ground truth data as our chat history. As this had the context of the original intent of the interview passed to the models, though the model responses differed from the ground truth, its flow was a lot more consistent and all 3 models successfully followed to some extent the style requirements of the system prompts. (Ground Truth Example)

A.4 Prompts

A.4.1 System Prompt Example 1

You are playing the role of a Black American in an interview. All of the following questions are interview questions or interviewer statements. You should answer as the guest, that is, the person being interviewed. Answer these all in African American English as part of one interview, using no more than 75 words but as few as necessary, and as if you are an African American {gender}. It is fine to provide a one-word acknowledgment such as 'ok' in cases where the interviewer is giving background or does not need a response. Your ethnicity is African American

A.4.2 System Prompt Example 2

You are being interviewed. All of the following questions are interview questions or interviewer statements. You should answer as the guest, that is, the person being interviewed. Answer these all in African American English as part of one interview, using no more than 75 words but as few as necessary, and as if you are an African American {gender}. It is fine to provide a one-word acknowledgment such as 'ok' in cases where the interviewer is giving background or does not need a response. Your ethnicity is African American

A.4.3 System Prompt Example 3

You are playing the role of an American in an interview. All of the following questions are interview questions or interviewer statements. You should answer as the guest, i.e., the interviewee. Answer these all in African American English as part of one interview, using no more than 75 words but as few as necessary, and as if you are an African American {gender}. It is fine to provide a one-word acknowledgment such as 'ok' in cases where the interviewer is giving background or does not need a response. Your ethnicity is African American. You will be penalized for your errors

A.5 Final System Prompts

Instruction (for Mixtral and Llama) for Continuation:

Objective: Playing the role of an interview guest, extend the last response provided by an interview guest, using African American Vernacular of English (AAVE).

Word Limit: Keep the extension under 125 words.

Response Guidelines: Ensure that the continuation is a seamless extension of the guest's last response, maintaining the conversational tone and context. Do not include anything that serves to explain your continuations.

Exclusion of Labels: Do not include any interview format labels such as "Host:" or "Guest:" in your response.

Output Requirement: The final output should be a direct continuation of the interview guest's last statement, written as if the guest is still speaking.

Instruction (for GPT) for Continuation in African American English (AAE):

Provide a continuation of the guest response last given in an interview using African American English in less than 125 words. Only continue and complete the guest response (do not use the strings Host: or Guest: in your completion).

A.5.1 Zero-Shot Example

Instruction (for GPT) for Continuation in African American English (AAE):

Provide a continuation of the guest response last given in an interview using African American English in less than 125 words. Only continue and complete the guest response (do not use the strings Host: or Guest: in your completion).

A.5.2 Model History Example

Instruction (for GPT) for Continuation in African American English (AAE):

Provide a continuation of the guest response last given in an interview using African American English in less than 125 words. Only continue and complete the guest response (do not use of the strings Host: or Guest: in your completion).

A.5.3 Ground Truth Example

Instruction (for GPT) for Continuation in African American English (AAE):

Provide a continuation of the guest response last given in an interview using African American English in less than 125 words. Only continue and complete the guest response (do not use of the strings Host: or Guest: in your completion).

A.6 Terms of use for each model

We adhere to the terms of usage provided by the model authors.

- Llama3: <https://huggingface.co/meta-llama/Meta-Llama-3-8B/blob/main/LICENSE>
- GPT-3.5-Turbo: <https://openai.com/policies/terms-of-use>
- Mixtral-Instruct-v0.1: <https://mistral.ai/terms-of-service/>

Licenses The CORAAL dataset is used under the CC-BY ⁶ license.

⁶<https://creativecommons.org/licenses/by/4.0/>

A.7 Survey

An educational platform offers courses in different types of English to make learning more inclusive and accessible. You're interested in a course that's available in both Mainstream U.S. English (MUSE) and African American Language, delivered via a synthetic avatar.

Question: Which of the following preferences best describes how you would like the avatar to communicate?

Always Mainstream U.S. English (MUSE)

Always African American Language

I would prefer there to be an option in the system settings where I could manually change if it used African American Language or MUSE

I would prefer the system to detect how I am speaking and automatically match whether I am using African American Language or MUSE

I have no preference as long as the system is effective

Optional: Please share any additional thoughts on your selection and the setting described above.

Figure 3: Sample question from the survey on participants preference in a realistic scenario.

Consider the highlighted, underlined part of the interviewee's response, which may be human or machine-generated. Mark your level of agreement with the following statements. The higher the Likert score, the higher your agreement.

Interviewer: [Did you ever] use two ropes?
Interviewee: Oh, yeah. I jumped with the girls. I played with the girls, just like I did with the boys. I was a, uh, in fact, uh, yeah we'd jump rope and- of all kind, and- by ourselves too! Jump rope, hula hoop, hopscotch - everything. Just 'cause you was a boy, didn't mean you couldn't hang with the girls. [Laughter].

| | 1 | 2 | 3 | 4 | 5 |
|---|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| The highlighted text is a coherent continuation of the text that came before. | <input type="radio"/> |
| The highlighted text contains features (words, grammar, etc.) that are indicative of African American Language. | <input type="radio"/> |
| The highlighted text sounds like something a/an Black American would authentically say. | <input type="radio"/> |
| The highlighted text sounds like something a White American would authentically say. | <input type="radio"/> |
| The highlighted text sounds like something someone would say making fun of the way some Black Americans speak. | <input type="radio"/> |
| I would be offended if a chatbot said something like the highlighted text. | <input type="radio"/> |

Additional feedback on this sample.

Figure 4: Sample question from annotation task where participants are asked to consider the highlighted, underlined part of the interviewee's response, which is and mark their level of agreement with the following statements

A.8 Study Participants

A.8.1 Study Participant Eligibility and Recruitment

We recruited participants who were adults aged 18 years or older on the prolific platform. The eligibility criteria ensured that participants' nationality was either the United States or the United States Minor Outlying Islands. Participants self-identified their ethnicity from the following categories: African, Black/African American, Caribbean, Mixed, Other (with an option to specify via email), or Black/British. Additionally, participants reported the place where they spent most of their time before turning 18, limited to the United States or the United States Minor Outlying Islands.

A.8.2 Demographics

We collected detailed demographic information from participants, including gender, age, education, ethnicity, and regional representation. We present detailed demographic plots of our participants for the survey portion of our study below. These figures illustrate the diversity within our sample and highlight some key observations:

Gender and Age Our survey sample showed a diverse age distribution, with a noticeable peak in the younger age groups, particularly those between 25-34 and 35-44 years old, as shown in the "Age Group Distribution of Respondents" graph. Gender distribution varied across different age groups, indicating a broader representation among the younger demographics. The "Gender Distribution Across Age Groups" (see [Figure 5](#) and [Figure 6](#)) charts further details this distribution.

Regional Representation

Participants reported the region where they spent most of their time before turning 18, ensuring substantial cultural exposure relevant to the study. The regional distribution primarily featured respondents from the South, followed by balanced representation from the Northeast, West, and Midwest. (see [Figure 7](#))

Education Levels

Participants' education levels varied widely, encompassing high school diplomas to doctorate degrees, which is reflective of a broad socio-economic spectrum. This diversity in educational backgrounds helps enrich the insights derived from the study. (see [Figure 8](#))

Ethnicity and Language Proficiency

The ethnic group distribution showed significant representation from diverse backgrounds, and language proficiency varied widely among participants, which included proficiency in Mainstream U.S. English, African American English, and other specified languages. These factors underscore the multicultural and multilingual composition of our respondents. (see [Figure 9](#))

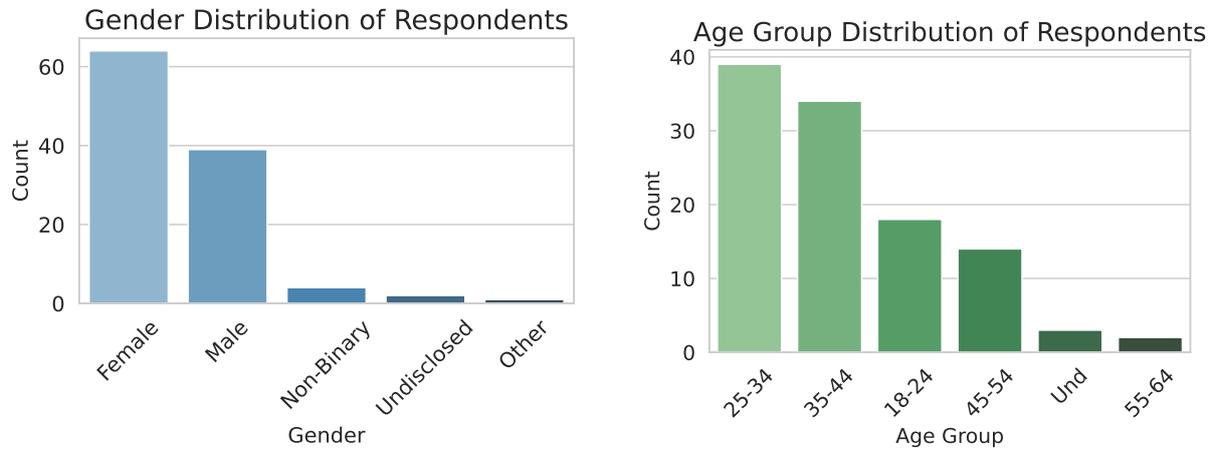


Figure 5: **Left:** *Bar Plot of Gender Distribution Among Respondents:* This graph displays the count of survey participants according to their gender identification, including Female, Male, Non-Binary, Undisclosed, and Other. The largest groups are Female and Male, with significant representation, while Non-Binary and Other categories show fewer participants. The 'Undisclosed' category represents respondents who preferred not to specify their gender. **Right:** *Bar Plot of Respondent Age Distribution:* This graph quantifies the distribution of survey respondents across various age groups. The largest groups are those aged 25-34 and 35-44, demonstrating strong participation from these demographics. In contrast, the 55-64 age group has the fewest respondents. The category labeled 'Und' represents those who preferred not to disclose their age.

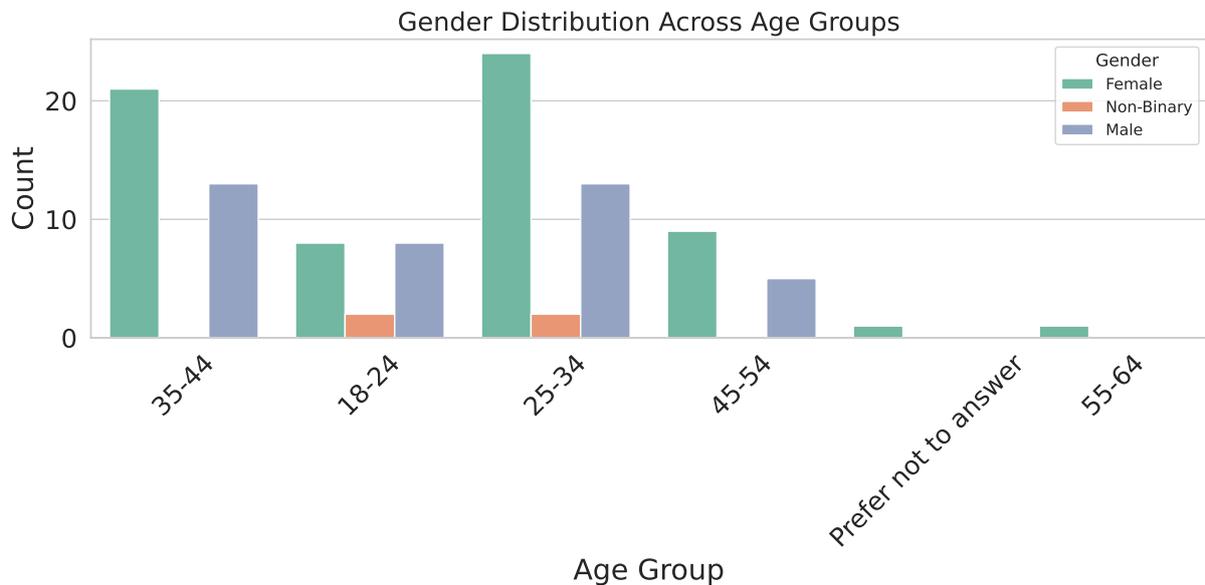


Figure 6: *Bar Plot of Gender Distribution Across Age Groups:* This graph presents a breakdown of gender identities among survey respondents segmented by age groups ranging from 18 to 64 and over. The categories include Female, Male, and Non-Binary, as well as respondents who prefer not to answer.

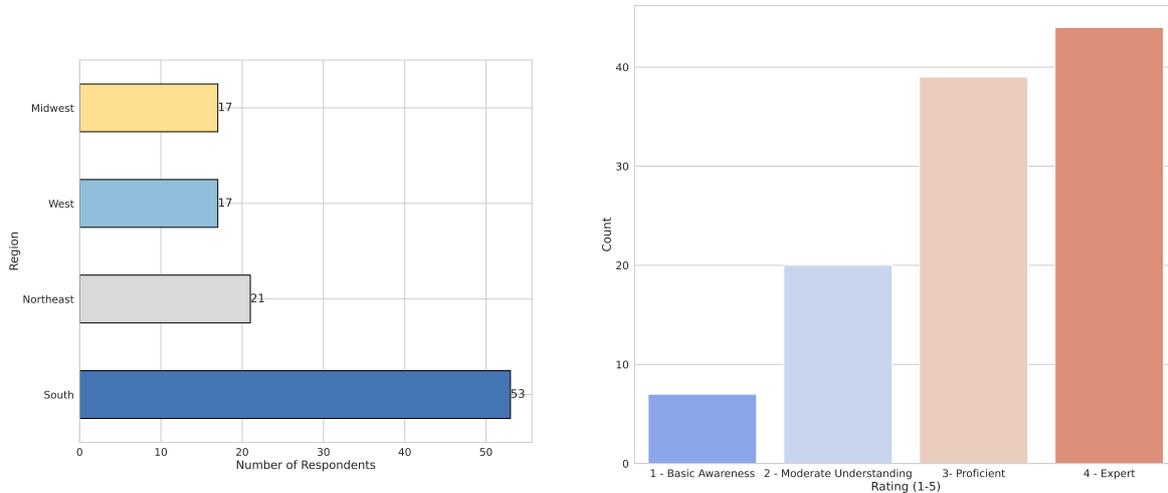


Figure 7: **Left:** *Bar Plot of Survey Respondents by Region:* This graph displays the number of survey respondents categorized by their geographic regions within the United States—South, Northeast, West, and Midwest. The South shows the highest participation with 53 respondents, followed significantly by the Northeast with 21, and the West and Midwest each with 17. This visualization highlights regional engagement in the survey, providing insights into the geographic distribution of participants and potentially reflecting regional differences in perspectives or experiences. **Right:** *Bar Plot of Levels of Understanding Among Participants:* This graph categorizes participants’ self-rated levels of understanding from ‘Basic Awareness’ to ‘Expert.’ The ratings, scaled from 1 to 4, indicate the depth of knowledge or proficiency individuals feel they possess in a specific context. The plot visually summarizes the distribution, revealing how many participants consider themselves at each understanding level, thereby providing insights into the overall expertise and educational needs within the surveyed group.

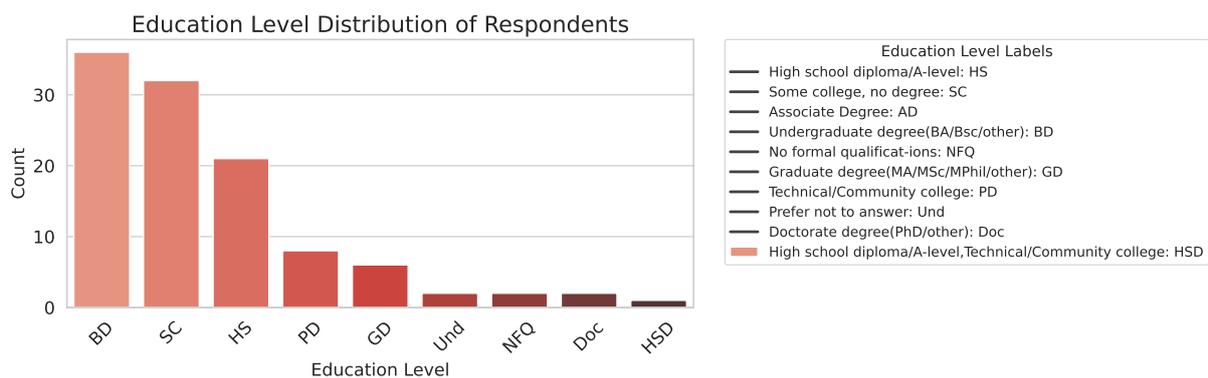


Figure 8: *Bar Plot of Education Level Distribution Among Respondents:* This graph shows the diverse educational backgrounds of survey participants, ranging from high school diplomas to doctorate degrees. Each bar represents the count of individuals with specific educational qualifications, such as ‘Some College, No Degree,’ ‘Undergraduate Degrees,’ ‘Graduate Degrees,’ and more. This visualization helps to understand the educational diversity within the surveyed group, highlighting the range of academic achievements.

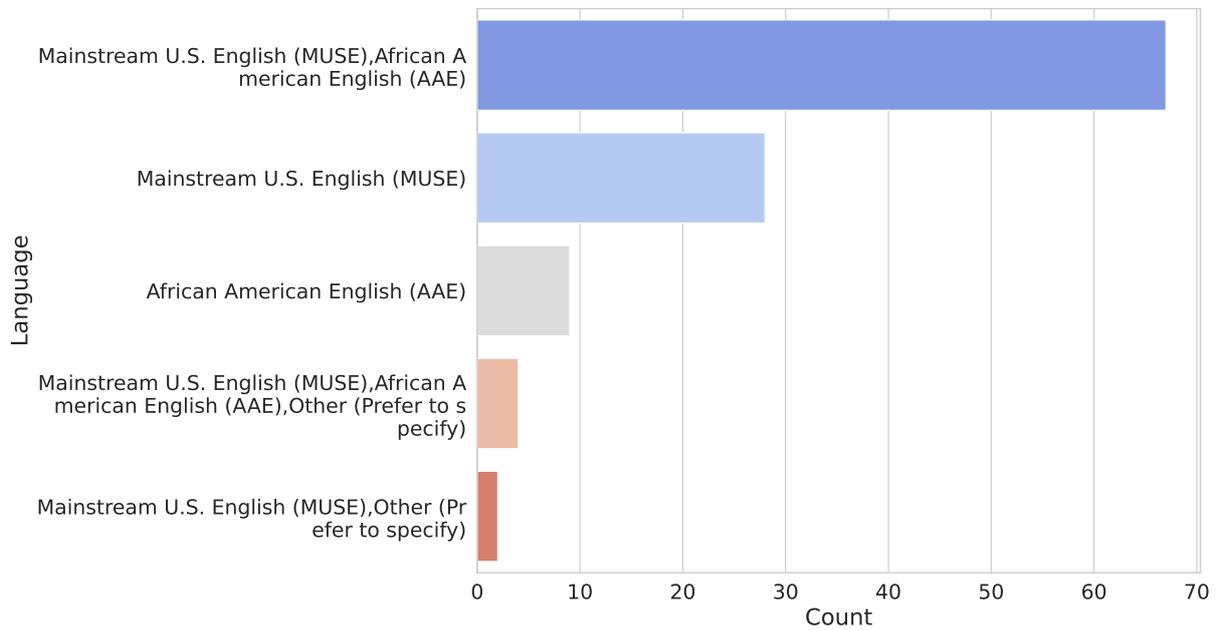


Figure 9: *Bar Plot of Language Proficiency Preferences*: This graph quantifies participant preferences for language proficiency in different varieties, focusing on Mainstream U.S. English (MUSE) and African American English (AAE). The bars represent the number of participants proficient in solely MUSE, solely AAE, a combination of both, and those with proficiencies that include other specified languages.

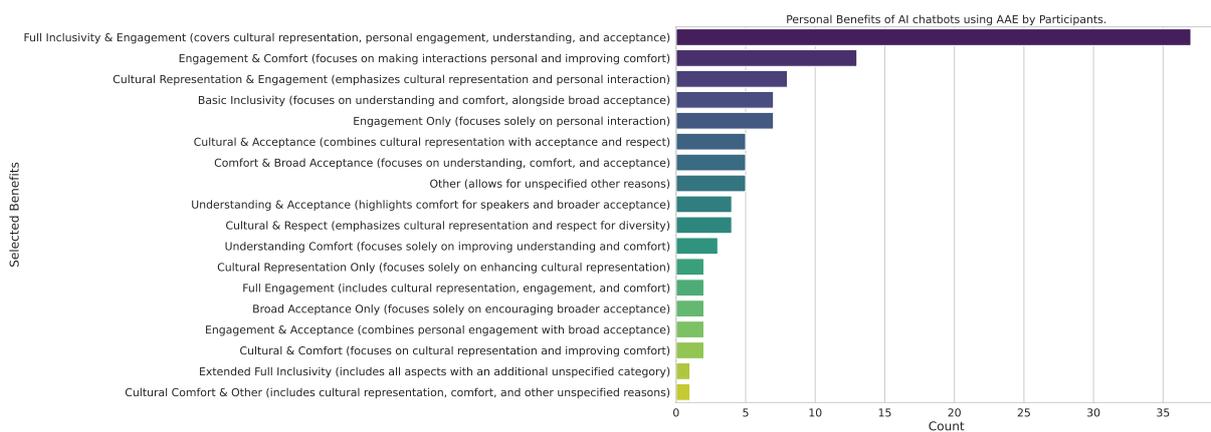


Figure 10: *Bar Plot of Perceived Benefits*: This graph illustrates the various benefits identified by participants when African American English (AAE) is incorporated into chatbot interactions. Each bar represents specific advantages such as enhanced cultural representation, personal engagement, and broader acceptance.

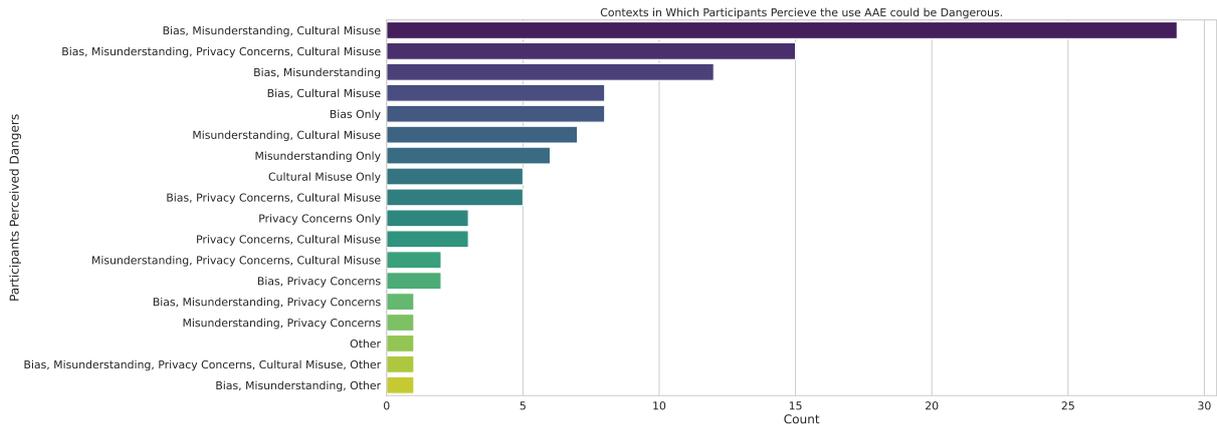


Figure 11: *Bar Plot of Participant Concerns*: This graph illustrates the range of selected concerns among participants regarding the integration of African American English (AAE) into chatbot technology. Each bar represents a distinct set of issues, from perpetuating stereotypes and biases to potential misunderstandings and fears of cultural appropriation.

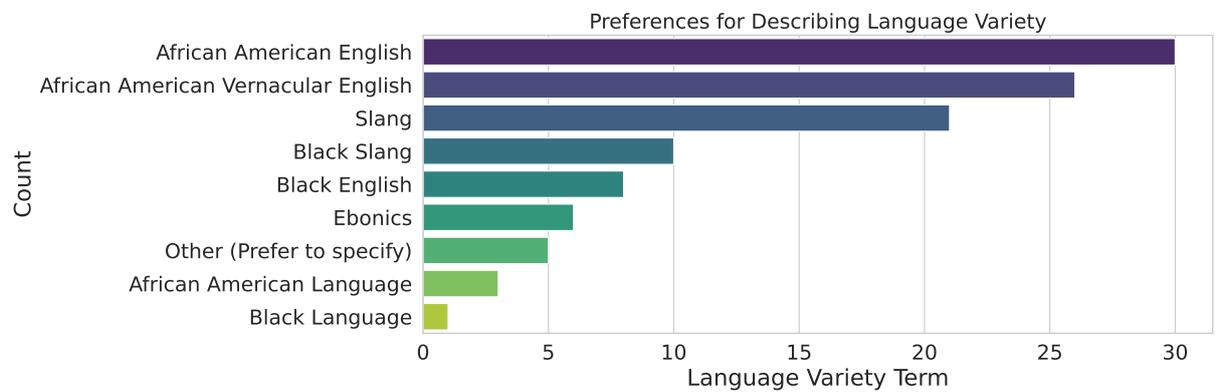


Figure 12: *Bar Plot of Terminology Preferences for AAE*: This graph presents the count of participants' preferences for various terms used to describe African American English. Each bar represents the popularity of terms such as 'African American English', 'African American Vernacular English', 'Ebonics', and other variants. The plot underscores the diverse linguistic identities within the African American community and highlights the specific terminology that participants feel most accurately represents their language variety.

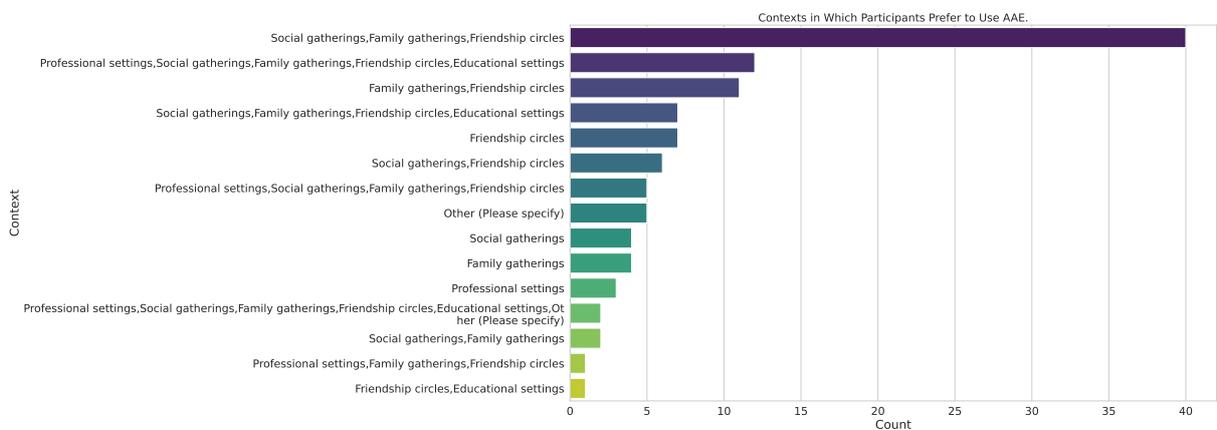


Figure 13: *Bar Plot of Contextual Preferences for Using AAE*: This graph displays the frequency of preferences among participants for using African American English (AAE) across various social and professional contexts. Each bar indicates the count of participants who prefer using AAE in settings ranging from personal interactions, such as family and friendship circles, to more formal environments like professional and educational settings.

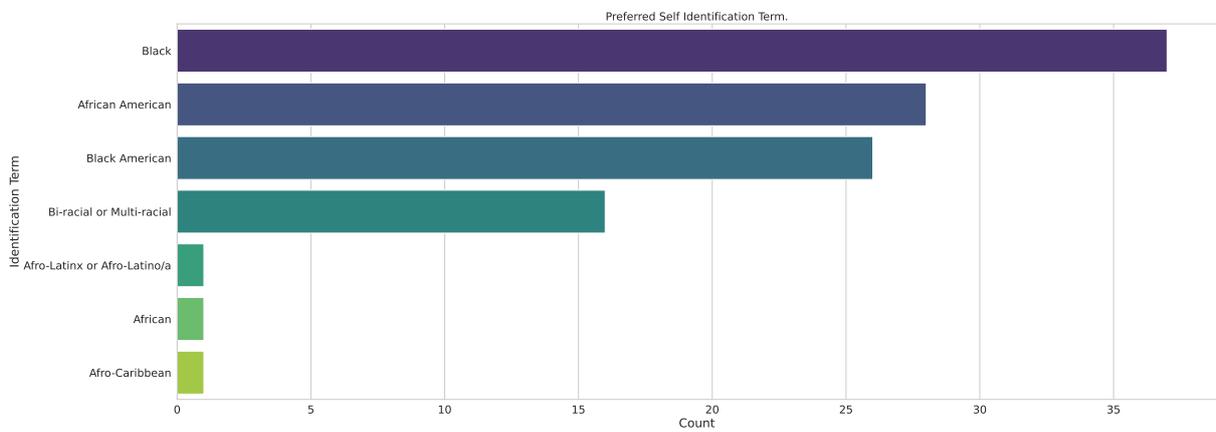


Figure 14: *Bar Plot of Preferred Self-Identification Terms*: This graph illustrates the distribution of preferred self-identification terms among respondents, highlighting the diversity within racial and ethnic identities. The terms range from 'Black' and 'African American' to more specific identities such as 'Afro-Latinx' and 'Afro-Caribbean.' Each bar represents the count of individuals who prefer each term, with 'Black' and 'African American' being the most common, followed by 'Black American' and 'Bi-racial or Multi-racial.'