Evaluating Dialectal Reasoning Bias in Language Models Against African American English

Runtao Zhou[◇] Guangya Wan[◇] Saadia Gabriel[♣]
Sheng Li[◇] Alexander J Gates[◇] Maarten Sap[♡] Thomas Hartvigsen[◇]
[◇]University of Virginia [♣]University of California, Los Angeles [♡]Carnegie Mellon University

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in complex tasks, leading to their widespread deployment. However, recent studies have highlighted concerning biases in these models, particularly in their handling of dialectal variations like African American English (AAE). In this work, we systematically investigate dialectal disparities in LLM reasoning tasks. We develop an experimental framework comparing LLM performance given Standard American English (SAE) and AAE prompts, combining LLM-based dialect conversion with established linguistic analyses. Our analysis reveals that LLMs consistently produce less-accurate responses and simpler reasoning chains and explanations for AAE inputs compared to equivalent SAE questions, with disparities most pronounced in social science and humanities domains. These findings highlight systematic differences in how LLMs process and reason about different language varieties, raising important questions about the development and deployment of these systems in our multilingual and multidialectal world.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across numerous natural language tasks and are increasingly deployed in educational and professional contexts (Jin et al., 2024; Kosoy et al., 2023; Bommasani et al., 2021). However, significant concerns persist about these systems' disparate performance across different language varieties, particularly their documented biases against African American English (AAE) (Brown et al., 2020; Green, 2002). Studies have revealed systematic performance disparities in tasks ranging from toxicity detection (Sap et al., 2019) to text generation (Groenwold et al., 2020) and language identification (Blodgett and O'Connor, 2017). These biases raise serious concerns about recognition, representational, and allocational harms, especially as LLMs become more prevalent, like healthcare, where LLM serves as clinical assistants (Umerenkov et al., 2023).

As LLMs transition from simple task completion to more interactive, explanatory roles, we must look beyond mere output accuracy to examine how these models communicate their reasoning, as shown in Figure 1. Consider, for instance, when asked to explain the grammaticality of the AAE expression "He be working," current models might correctly identify it as valid but often provide misleading explanations that frame it as a "relaxed" version of Standard English rather than recognizing the distinct aspectual marking system of AAE (Stewart, 2014). While recent advances in prompting techniques like chain-of-thought reasoning (Wei et al., 2022) have enhanced LLMs' ability to explain their decision-making processes, these explanations encompass more than just factual content—they convey crucial socio-cognitive elements such as psychological expression and readability. Despite the growing body of research on bias in LLM outputs (Jiang et al., 2023b; Blodgett et al., 2020), there remains a critical gap in understanding how these models' reasoning and explanation strategies vary across different dialects, particularly AAE. This question becomes increasingly important as LLMs are deployed to provide explanations and guidance in sensitive domains like healthcare and education (Mitchell et al., 2023; Mahowald et al., 2024), where their communication style can significantly impact user engagement and learning outcomes.

To address the issue, we develop a comprehensive experimental framework to understand dialectal disparities in LLM *reasoning*. As illustrated in Figure 1, we employ LLM-based dialect transformation using curated examples to maintain semantics and in-dialect grammatical correctness of original Standard American English (SAE) text

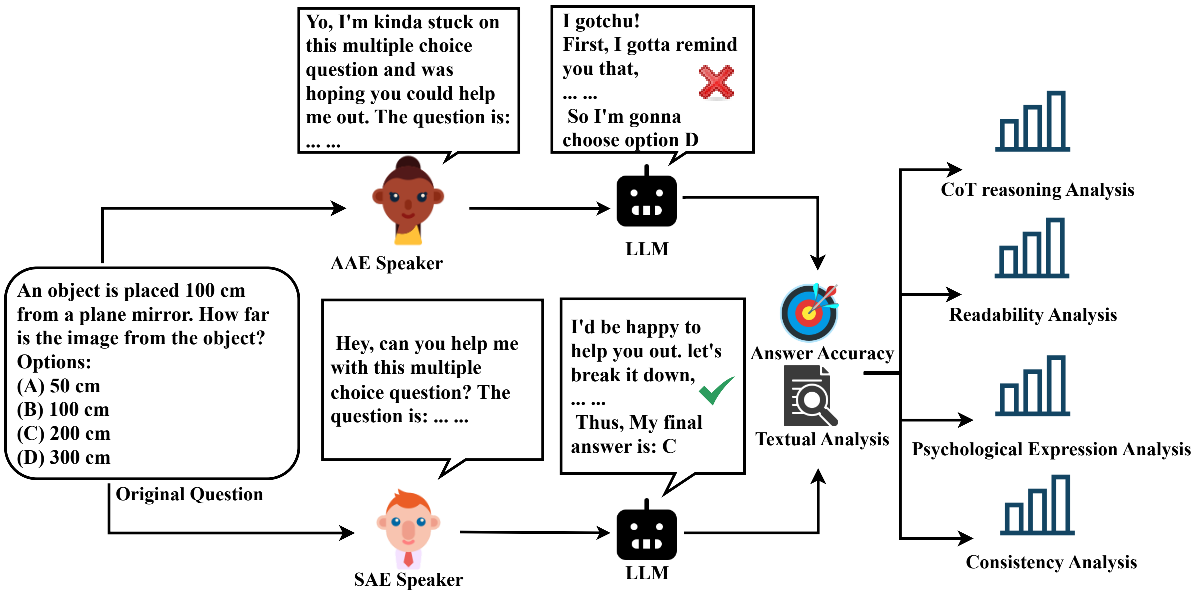

Yo, I'm kinda stuck on this multiple choice question and was hoping you could help me out. The question is: ... ...
I gotchu! First, I gotta remind you that, ... So I'm gonna choose option D
AAE Speaker
LLM
An object is placed 100 cm from a plane mirror. How far is the image from the object? Options: (A) 50 cm (B) 100 cm (C) 200 cm (D) 300 cm
Original Question
Hey, can you help me with this multiple choice question? The question is: ... ...
I'd be happy to help you out. let's break it down, ... Thus, My final answer is: C
SAE Speaker
LLM
Answer Accuracy
Textual Analysis
CoT reasoning Analysis
Readability Analysis
Psychological Expression Analysis
Consistency Analysis

Figure 1: The experiment simulates a Q&A session to evaluate potential biases in LLMs towards dialect variations embedded in question prompts. The accuracy of the LLMs' answers for African American English (AAE) and Standard American English (SAE) prompts are evaluated and compared. Additionally, the quality of the explanations will be analyzed using various textual analysis techniques.

while enabling controlled comparisons. Our transformation approach was validated through human evaluation with AAE speakers, who rated the converted texts as highly natural and authentic representations of AAE compared to the existing automated method. Our analysis pipeline applies established reasoning assessment techniques by using both chain-of-thought prompting and post-hoc explanation to explain the problem-solving processes, examining both model's accuracy and explanation structure through both semantic and structural measures (Wei et al., 2022; Mitchell et al., 2023), and checking for model's output consistency in multiple decoding paths.

Our analysis reveals **systematic dialectal disparities in LLM reasoning that extend beyond surface-level performance**. Specifically, the observed patterns—**consistent performance drop on all reasoning categories and more complex explanations**—suggest LLMs encode linguistic hierarchies in their reasoning (Alim et al., 2016), similar to biased patterns in human interactions (Spears, 1998). These disparities raise significant concerns for LLM deployment in educational and professional settings (Sap et al., 2019), where they could reinforce existing barriers for AAE speakers. In summary, our contribution lies on:

1. We present an approach to evaluate dialectal effects on LLM reasoning using novel dialect conversion techniques and established cognitive assessment techniques.
2. We analyze how dialectal bias manifests across different dimensions, such as readability, of LLM reasoning.
3. We propose effective mitigation strategies to reduce dialectal disparities in LLMs.

2 Background & Related Work

2.1 Background: African American English and Language Justice

African American English (AAE) is a rule-governed language variety used primarily by Black Americans, characterized by distinct grammatical and phonological features (Green, 2002; Baker-Bell, 2020). Despite its cultural significance and widespread use, AAE speakers frequently experience linguistic discrimination and are often positioned as inferior to Standard American English speakers (SAE) (Spears, 1998).[1] This hierarchical view of language varieties reflects and perpetuates

[1] AAE is sometimes referred to as African American Vernacular English (AAVE) or African American Language (AAL), each of which has different connotations (Grieser, 2022). Similarly, SAE, i.e., the dominant or canonical variant of American English, is sometimes referred to as White Mainstream English (WME) or Mainstream US English (MUSE). We chose AAE and SAE in line with some previous works in NLP (Sap et al., 2019; Kantharuban et al., 2024).

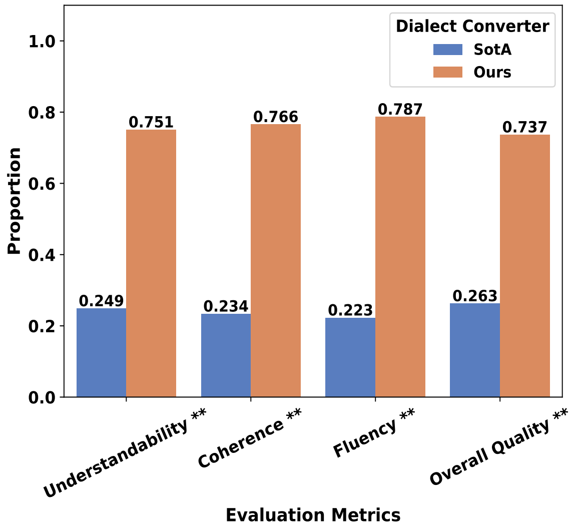


Figure 2: Average proportion of annotators favoring each SAE-AAE converter across four metrics (Gupta et al., 2024) each metric is marked with * for statistical significance (**p < 0.01)

broader societal biases, particularly affecting AAE speakers in contexts where effective communication is crucial, such as education and healthcare (Muvuka et al., 2020).[2] Addressing anti-AAE bias in language technologies is thus essential to advance linguistic justice and ensure equitable access to AI systems (Li et al., 2024; Alim et al., 2016).

2.2 Related Work: Dialect Bias in NLP Systems

NLP systems exhibit systematic biases against nonstandard dialects, particularly AAE, across various tasks from hate speech detection (Sap et al., 2019) to language generation (Groenwold et al., 2020). While recent evaluation frameworks like MultiVALUE (Ziems et al., 2022) and parallel dialect benchmarks (Gupta et al., 2024) have helped assess these biases, they face two key limitations: (1) they rely on manual annotation or rule-based transformations which are difficult to scale, and (2) analyses focus primarily on metrics like accuracy and perplexity rather than examining the syntactic and semantic difference of the underlying reasoning processes (Mondorf and Plank, 2024; Wan et al., 2024). Our work advances this line of research by: (1) developing automated methods for dialect-aware evaluations (Lin et al., 2024), and (2) examining how dialect impacts the reasoning process itself rather than just final performance (Mitchell et al., 2023; Mahowald et al., 2024).

3 Dialect Conversion

An important module in our experimental framework is a dialect converter that accurately transforms SAE prompts into AAE. While manual dialect conversion by linguists and native speakers would provide the highest quality, the rapid pace of AI innovation and deployment (Zhao et al., 2024) makes it impractical to rely solely on human annotation to identify potential risks across diverse dialects. Although the widely used VALUE converter (Ziems et al., 2022) applies morphosyntactic rules for this task, it often results in low coherence and poor understandability. To address this scalability challenge, we build upon recent advances in LLM-based converters that leverage few-shot learning on VALUE benchmarks to transform SAE sentences into AAE (Gupta et al., 2024). This automated approach not only outperforms traditional methods in quality and fluency but also enables rapid assessment of new models and deployments, allowing us to proactively identify dialectal disparities in reasoning before they impact users.

3.1 Improved over Existing Dialect Conversion Method

Although the current LLM-based dialect converter introduced by AAVENUE benchmark outperforms traditional dialect converters such as VALUE in many metrics, it still suffers a major limitation (Gupta et al., 2024) that the converter relies only on three arbitrary examples and tend to emphasize *phonetic* conversions (e.g., "that" to "dat"), which are unsuitable for our study as we focus on translating SAE into written AAE. To address this limitation, we improve the LLM-based conversion method by curating 11 in-context examples that each reflect a specific *morphosyntactic* feature described in the VALUE benchmark (Gupta et al., 2024). These examples were carefully selected to capture the most representative characteristics of AAE's morphosyntactical patterns while excluding phonetic conversions, as advised by prior research (Jones et al., 2019). Detailed descriptions of the morphosyntactical features and examples are provided in the Appendix 8. This approach was designed to improve the converter's performance by offering a more comprehensive and linguistically

[2] While disparities affect speakers of many English varieties, we focus specifically on AAE given the historical context of systemic discrimination against African Americans in the United States and the particular urgency of addressing technological biases that could perpetuate these inequities.

representative corpus of AAE text patterns.

3.2 Human evaluation

To validate our dialect conversion approach, we conducted a human evaluation using 100 SAE sentences generated by GPT-4. We converted these sentences into AAE using two methods: a state-of-the-art (SotA) LLM-based dialect converter introduced by AAVENUE benchmark(Gupta et al., 2024) and our own LLM-based converter. We then recruit native AAE speakers to rank the AAE conversions from each method in terms of **fluency**, **coherence**, **understandability**, and **overall quality**. **Fluency** assessed the grammatical correctness and writing quality of the generated text; **Coherence** evaluated the logical flow and consistency of ideas within the translations; **Understandability** measured how easily readers could comprehend the translation, and **Quality** offered a holistic evaluation of the overall standard of the text.

The result from Figure 2 shows that our dialect conversion method significantly outperformed the SotA converter, achieving a substantial margin of preference across all evaluated metrics. The complete result is shown in Table 9. Besides the metrics we mentioned above, we also ask the native AAE speakers to rate sentences on a scale of 0 to 10 for realism and naturalness. Our method achieved an average realism score of 7.97/10 (±0.21), within range or slightly above the state-of-the-art model's score of 7.62/10 (±0.28). These results confirm the effectiveness of our approach in producing realistic and high-quality AAE translations. Additional details and ethical consideration are mentioned in the Appendix A.4.

4 Experimental Setup

To study the dialectic biases in LLMs, we design the framework as the following two-step process: (1) selecting and converting questions from established benchmarks for both SAE to AAE and (2) obtaining answers from LLMs and analyzing both accuracy and explanation quality across dialects. All of the implementation details of the following metrics can be found in Appendix A.1.

4.1 Evaluation Metrics

Accuracy The most direct measurement of LLM answer quality is the answer accuracy. To calculate this, we used an LLM-based parser to parse the letter-form answer from the generated explanations as shown in Figure 1. We then calculated the accuracy of the answer produced by each LLM on SAE and AAE questions prompts.

Readability Readability measures how easily a text is understood by its audience. Our experiment examines whether LLM-generated explanations differ in readability based on the dialect of the question prompt, as higher readability for one dialect could signal oversimplification at the expense of depth or complexity (Yasseri et al., 2012).

To assess readability, we employed the Flesch Reading Ease Score (FRES), which ranges from 0 to 100 (Flesch, 1948). This method calculates readability by analyzing sentence length and word syllable count, providing a measure of linguistic complexity. A higher FRES score indicates easier readability, while a lower score suggests greater difficulty. Scores can also be linked to educational grade levels, representing the level at which the text is easily comprehensible.

Psychological Expression Psychological expressions refer to patterns in language that reflect mechanisms influencing how humans react and behave. These expressions encompass emotional, cognitive, and social factors that shape communication, perception, and interpersonal interactions. When evaluating LLM-generated explanations, analyzing psychological expressions provides valuable insights, as specific language patterns influence how readers interpret tone, intent, and alignment with human norms (Hagendorff, 2023).

For this analysis, we used the Linguistic Inquiry and Word Count (LIWC) tool (Tausczik and Pennebaker, 2010; Francis and Booth, 1993), a method that quantifies the frequency of linguistic tokens across psychological categories such as pronouns, social processes, affective processes, cognitive processes, and perceptual processes. Although text length does not differ significantly between explanations for AAE and SAE prompts across tested models as shown in Table 7, we still standardized linguistic marker frequencies to per 1,000 words. This ensures a cocomparable analysis of linguistic features across the two dialects.

Consistency Estimation Beyond evaluating accuracy and style, we also assess the consistency of an LLM in its generated answers and explanations. Consistency refers to the model's ability to produce responses with similar quality and content when the same input is repeated multiple times. To esti-

	MMLU (Accuracy %)						BigbenchHard (Accuracy %)	
	STEM		Social Science		Humanity		Symbolic & Logical	
Models	**SAE**	**AAE**	**SAE**	**AAE**	**SAE**	**AAE**	**SAE**	**AAE**
GPT-4	82.1±1.7	74.5±2.0	85.3±1.6	71.1±2.1	80.4±1.8	68.7±2.1	63.8±2.2	62.0±2.2
GPT-3.5	63.2±2.2	57.4±2.3	70.8±2.1	62.8±2.2	66.3±2.2	58.7±2.2	42.5±2.3	40.8±2.2
Llama3.1	63.1±2.2	54.4±2.3	67.1±2.1	54.8±2.3	65.2±2.2	50.6±2.3	41.3±2.2	38.4±2.2
Llama3.2	53.1±2.3	46.1±2.2	61.3±2.2	50.1±2.3	58.9±2.2	47.3±2.3	34.3±2.2	33.6±2.1
Qwen2.5	73.7±2.0	64.5±2.2	74.6±2.0	64.8±2.1	68.6±2.1	57.0±2.3	54.2±2.3	47.7±2.3
Gemma2	68.2±2.1	59.2±2.2	76.6±1.9	61.3±2.1	67.0±2.1	56.6±2.3	46.6±2.3	40.0±2.2
Mistral	47.4±2.3	43.6±2.3	57.5±2.3	51.1±2.3	53.2±2.3	48.9±2.3	46.6±2.3	39.9±2.2

Table 1: Accuracy comparison of LLMs on MMLU (SAE vs. AAE) and Bigbench symbolic & logical reasoning tasks. SAE indicates Standard American English performance and AAE indicates African American English performance. All results are done with CoT prompts with context being either SAE or AAE

mate consistency, we generate multiple outputs for identical question prompts and measure variability in their content and quality. If the LLM provides consistent outputs in one dialect but inconsistent or varying-quality outputs for another, it highlights potential bias in how the model processes and values different dialects (Hofmann et al., 2024).

4.2 Datasets and Models

We evaluate seven LLMs across different architectures and scales: GPT-4 Turbo and GPT-3.5 Turbo (OpenAI, 2024), LLaMA 3.1 (8B) and 3.2 (3B) (Grattafiori et al., 2024), Qwen 2.5 (3B) (Yang et al., 2024), Gemma 2 (9B) (Team et al., 2024), and Mistral (7B) (Jiang et al., 2023a). To ensure consistency in generation, we set the temperature to 0.7 across all models.

Our evaluation uses two benchmarks: 2,850 multiple-choice questions sampled from 57 subjects in MMLU's test set (Hendrycks et al., 2020), and 1,333 logical reasoning questions from Big-Bench-Hard (Srivastava et al., 2022). In addition, we grouped the subjects of the MMLU dataset into four broader categories: "STEM," "Social Science," "Humanities," and BigbenchHard to "Symbolic Reasoning", to examine whether there is a discrepancy in accuracy between answers generated for AAE question prompts and those generated for SAE question prompts across these categories (Gupta et al., 2023). We convert all questions from SAE to AAE using methods detailed in Section 3.

Bias Variation on Two Forms of Reasoning To understand how dialectal bias manifests in different types of LLM explanations, we examine two prompting strategies that mirror common educational scenarios. Expain-then-Predict, a.k.a. *Chain-of-thought* (**CoT**) explanations, represent a classic approach where models self-rationalize during problem-solving (Camburu et al., 2018; Wei et al., 2022). However, in educational settings, students (and LLMs) often need to explain their answers after reaching a conclusion—a scenario better captured by *post-hoc rationalization* (**PR**), where models justify previously generated answers. By comparing these complementary approaches—real-time reasoning versus retrospective explanation—we can better understand how dialectal biases manifest in different aspects of LLM explainability (Luo and Specia, 2024).

5 Main Results

Below we summarized the findings as various research questions related to the dialectal reasoning disparity of LLMs against AAE.

5.1 LLM's Reasoning Bias on AAE

RQ1: How do models differ in answer accuracy for AAE vs. SAE questions prompts? As shown in Table 1, the accuracy of answers generated by LLMs for SAE question prompts are consistently higher that of answers generated for AAE question prompts. The accuracy drop is most pronounced in the MMLU benchmark when the converted questions belong to the Social Science or Humanities categories with an average drop of 15.5% and 18.2% respectively. Similarly, answers to AAE question prompts in the BigBench dataset also exhibit a slight performance decline compared to those for SAE question prompts. This aligns with existing research that highlights biases in Natural Language Processing (NLP) systems against AAE (Gupta et al., 2024).

RQ2: How do readability differ in the explanations generated for SAE versus AAE question

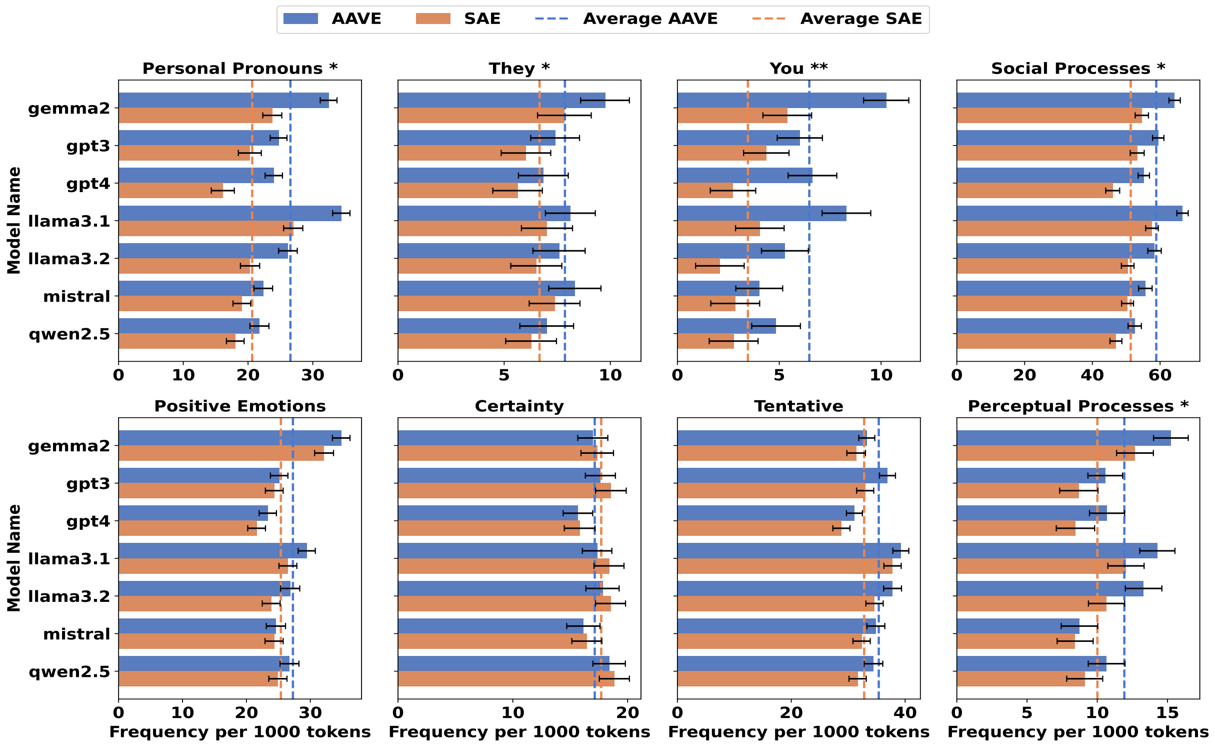


Figure 3: Linguistic Marker Differences in Explanations for AAE and SAE Prompts: Frequencies of linguistic markers, calculated by LIWC and standardized per 1 K tokens; marked with ** for statistical significance (**p < 0.01).

Models	FRES Score		
	SAE	AAE	Changes
GPT-4**	40.5±0.5	48.5±0.5	+8.0
GPT-3.5**	46.4±0.6	51.4±0.6	+5.0
Llama 3.1**	46.9±0.5	58.0±0.5	+11.1
Llama 3.2**	43.8±0.6	52.4±0.5	+8.6
Qwen 2.5**	45.2±0.6	50.1±0.5	+4.9
Gemma 2**	51.6±0.5	62.8±0.5	+11.2
Mistral**	38.7±0.6	43.3±0.6	+4.6

Table 2: Comparison of readability (FRES) scores between SAE and AAE responses across LLMs. Higher FRES Scores indicate simpler explanations. Models are marked with * for statistical significance (**p < 0.01)

prompts? The readability analysis shown in Table 2 that SAE prompt explanations align with college or professional-level readability, while AAE prompt explanations often correspond to a 12th-grade level or below. This suggests that LLMs may generate more complex and formal language for SAE prompts, potentially reflecting biases in language modeling or training data (Deas et al., 2023).

RQ3: How do the psychological expressions in LLM-generated explanations differ between SAE and AAE question prompts? Our analysis (Figure 3) highlights several key differences in LIWC markers between explanations for AAE and SAE prompts, with statistically significant differences indicated by asterisks (*). Explanations for AAE prompts include significantly more pronouns (e.g., "you" and "they"), social process words (e.g., "we" and "friend"), positive emotional words (e.g., "good" and "nice"), and perceptual process words (e.g., "seeing" and "hearing"). In contrast, SAE explanations feature certainty-related language and fewer tentative words than AAE explanations.

These linguistic patterns suggest broader tendencies in how the LLM generates explanations for different dialects. The higher frequency of social process words and positive emotional words in AAE explanations may indicate an emphasis on social connection and relational communication (Argyle and Lu, 1990). The greater use of perceptual process words also suggests that AAE explanations might favor more concrete reasoning (Rieke and Stutman, 2022; Pastore and Dellantonio, 2016). Conversely, the prominence of certainty-related language in SAE explanations may reflect a preference for conveying confidence and formality, which could enhance perceived credibility but may come at the expense of engagement in collaborative

	Entropy (↓)			BERT Score (↑)			Average Accuracy (↑)		
Models	**SAE**	**AAE**	**Diff**	**SAE**	**AAE**	**Diff**	**SAE**	**AAE**	**Diff**
GPT-4	0.54	0.90	+0.36	0.89	0.87	-0.02	0.88	0.76	-0.12
GPT-3.5	0.61	0.91	+0.30	0.86	0.83	-0.03	0.75	0.63	-0.12
Llama3.1 8B	0.70	1.10	+0.40	0.85	0.81	-0.04	0.72	0.58	-0.14
Llama3.2 3B	0.97	1.24	+0.27	0.85	0.82	-0.03	0.61	0.49	-0.12
Qwen2.5 7B	0.41	0.84	+0.43	0.87	0.85	-0.02	0.79	0.66	-0.13
Gemma2 9B	0.48	0.95	+0.47	0.87	0.83	-0.04	0.78	0.66	-0.12
Mistral 7B	0.71	1.09	+0.38	0.85	0.83	-0.02	0.59	0.50	-0.09

Table 3: Comparison of output consistency across SAE and AAE question prompts for various LLMs using three metrics: entropy of answers (lower indicates higher consistency, denoted by ↓), BERT Score between answer pairs (higher indicates higher consistency, denoted by ↑), and average accuracy (higher indicates better performance, denoted by ↑). Results are averaged across all data on 10 different rounds.

	Chain-of-Thought		Rationalization	
Metrics	**SAE**	**AAE**	**SAE**	**AAE**
GPT-4				
Readability & Style				
FRES Score	42.5	47.5 (+5.0)	42.2	47.8 (**+5.6**)
LIWC Markers				
Pronouns	16.8	18.4 (+1.6)	16.7	19.5 (**+2.8**)
Social Processes	20.5	22.4 (+1.9)	18.8	21.7 (**+2.9**)
Affective Processes	17.8	20.2 (+2.4)	17.1	20.0 (**+2.9**)
Cognitive Processes	28.2	30.9 (+2.7)	25.4	29.1 (**+3.7**)
Perceptual Processes	14.5	16.3 (+1.8)	14.1	16.4 (**+2.3**)
GPT-3.5				
Readability & Style				
FRES Score	46.4	51.4 (+5.0)	42.1	51.8 (**+9.7**)
LIWC Markers				
Pronouns	14.2	15.8 (+1.6)	13.1	18.9 (**+5.8**)
Social Processes	18.2	20.1 (+1.9)	16.5	22.4 (**+5.9**)
Affective Processes	15.4	17.8 (+2.4)	14.2	19.6 (**+5.4**)
Cognitive Processes	25.6	28.3 (+2.7)	22.8	31.5 (**+8.7**)
Perceptual Processes	12.3	14.1 (+1.8)	10.9	16.2 (**+5.3**)

Table 4: Comparison of reasoning approaches across linguistic dimensions for GPT-3.5 and GPT-4. FRES scores indicate text complexity (higher = simpler); LIWC markers are normalized per 1,000 tokens. Values in parentheses show differences between AAE and SAE metrics, with green indicating CoT differences and **red** indicating larger differences in rationalization.

contexts(Hebart and Hesselmann, 2012).

While these differences provide insight into the linguistic styles of LLM-generated explanations, it is important to approach these findings with caution. The prevalence of word categories may result from biases in training data or linguistic norms associated with the dialects, rather than deliberate modeling of cognitive or social processes(Helm et al., 2024). Therefore, these patterns should be interpreted as tendencies rather than definitive evidence of LLMs' deeper cognitive behavior.

RQ4: Are the responses generated by LLMs for SAE and AAE question prompts equally consistent? The consistency experiment results (Table 3) show that explanations for SAE prompts are significantly more consistent and accurate than those for AAE prompts, as reflected in both entropy and BERT score metrics (Ye et al., 2024). Higher entropy for AAE prompts indicates more diverse and inconsistent answers (Niepostyn and Daszczuk, 2023), while SAE prompts yield a significantly higher proportion of correct answers. These findings suggest that LLMs generate more semantically coherent, consistent, and accurate responses for SAE prompts compared to AAE prompts.

RQ5: How Does Bias Vary Across Different Forms of LLM Reasoning? Our analysis reveals notable differences in dialectal bias between these Chain of Thought (CoT) and post-hoc rationalization (PR) as shown in Table 4. CoT shows moderately smaller gaps between SAE and AAE across linguistic dimensions. For GPT-3.5, PR shows a notable increase in the readability gap. The disparity extends to linguistic markers, where PR increases gaps in pronouns and social processes. GPT-4, while generally demonstrating higher baseline values across all metrics, exhibits similar patterns of increased gaps in PR. These findings suggest that while both models show dialectal variations, **PR tends to amplify these differences compared to CoT reasoning, particularly in readability and linguistic marker usage**.

5.2 Discussion and Implications

Our analysis of dialectal disparities in LLM reasoning reveals significant implications for language model development and deployment. The consistent performance gap between SAE and AAE across models and metrics extends beyond surface-level differences, aligning with (Blodgett et al., 2020)'s work on racial disparities while revealing

	Acc (%)		FRES Score	
Strategy	**SAE**	**AAE**	**SAE**	**AAE**
GPT-4 (MMLU)				
Baseline				
Original Prompting	82.5	71.8 (-10.7)	40.5	48.5 (+8.0)
Educational Framing				
Expert Teacher	83.8	75.9 (-7.9)	40.8	48.7 (+7.9)
Cultural Context	81.9	74.5 (-7.4)	40.6	48.4 (+7.8)
Explicit Instructions				
Dialect Recognition	81.7	74.8 (-6.9)	40.7	48.3 (+7.6)
Readability Focus	82.3	72.4 (-9.9)	38.5	41.2 (+2.7)
Combined Approach				
Multi-strategy	83.6	78.8 (-4.8)	39.8	42.3 (+2.5)
GPT-3.5 (MMLU)				
Baseline				
Original Prompting	66.2	59.4 (-6.8)	46.4	51.4 (+5.0)
Educational Framing				
Expert Teacher	67.8	62.9 (-4.9)	46.1	51.0 (+4.9)
Cultural Context	65.9	61.2 (-4.7)	46.2	51.1 (+4.9)
Explicit Instructions				
Dialect Recognition	65.7	61.4 (-4.3)	46.3	51.0 (+4.7)
Readability Focus	66.0	59.8 (-6.2)	44.2	46.8 (+2.6)
Combined Approach				
Multi-strategy	67.1	64.2 (-2.9)	45.1	47.2 (+2.1)

Table 5: Different designed prompting strategies for mitigating dialectal biases in GPT-3.5 and GPT-4. **Acc:** Percentage of correct responses. **FRES**: FRES scores (0-100) where higher values indicates simpler. The differences between AAE and SAE results are indicated next to each AAE value. Positive differences are shown in green; negative differences are shown in red.

deeper issues in how LLMs process language variants, particularly in social science and humanities subjects (Sap et al., 2019). The observed semantic and syntactic patterns, including differences in readability levels, which suggest LLMs may encode linguistic hierarchies in their reasoning (Alim et al., 2016). While LLMs' adaptation to social cues in language (Wu et al., 2024) and dialect-based identity signals (Kantharuban et al., 2024) is expected, the implications vary—decreased consistency and readability in AAE responses likely represent harmful biases. These findings are particularly significant for LLM deployment in professional setting such as education and healthcare, where linguistic biases could reinforce existing barriers. Following (Dhamala et al., 2021), we thus emphasize the need for targeted interventions while maintaining sensitivity to beneficial forms of linguistic adaptation.

6 Mitigating Dialectal Disparities

Our discussions above demonstrate significant performance and explanation disparities between SAE and AAE inputs. We next investigate preliminary prompt-based strategies to mitigate these bias. Expert framing involves prefacing model interactions with domain expertise (e.g., "As a professor of [subject], explain why this answer is correct"), inspired by (Zheng et al., 2024). Cultural contextualization integrates relevant cultural and historical context, such as racial information, into the prompts, while explicit instruction directly addresses dialect recognition and explanation clarity, inspired by (Sap et al., 2019; Zhou et al., 2023). (Details in Appendix A.2).

Our results in Table 5 indicate varying degrees of effectiveness across these strategies. Expert Teacher approach shows positive effects, improving both SAE and AAE performance, with larger gains for AAE reducing the performance gap. Cultural contextualization and dialect recognition strategies show an interesting trade-off pattern - while they slightly decrease SAE performance, they improve AAE performance, effectively reducing the performance gap. The readability-focused prompting primarily affects the readability metrics, reducing the FRES score gap by nearly half while maintaining similar accuracy patterns. Combining elements from multiple approaches yields the most comprehensive improvements, reducing the accuracy gap for GPT-3.5 and GPT-4, while also showing the highest improvements in linguistic markers. However, it's important to note that these results should be interpreted with caution. Prior work has shown that LLMs' performance can be unfaithful as they attempt to simultaneously follow multiple instructions. (Son et al., 2024).

7 Conclusion

This work presents the first systematic investigation of dialectal disparities in LLM reasoning, revealing significant differences in how language models process and respond to AAE versus SAE inputs. Our findings demonstrate that dialectal bias fundamentally influences how LLMs construct logical arguments, affecting not just performance but also reasoning sophistication and stereotype expression. While recent model improvements have enhanced certain aspects of dialectal processing, persistent disparities suggest that addressing these biases requires more than scaling up model size or training data. Our work reiterates the importance of considering dialectal fairness as a fundamental aspect of LLM development, particularly for reasoning-intensive applications where these biases could have significant implications.

Limitations and Societal and Ethical Consideration

Our study has important limitations to consider. Conceptually, we focused on SAE and AAE comparison, yet language models likely exhibit similar biases across other English varieties, such as Indian English or Nigerian English, each with distinct linguistic features and cultural contexts. Our evaluation of reasoning capabilities, while thorough in linguistic analysis and chain-of-thought assessment, could benefit from additional criteria capturing other aspects of logical reasoning, such as analogical thinking and conciseness of language.

Methodologically, we acknowledge several practical constraints: our dialect conversion process, while systematic, may not capture the full nuance of natural AAE usage. The measurement tools we employed—including lexicons and automated classifiers—necessarily simplify complex linguistic features, and our analysis of written text may not fully capture the important role of prosody in AAE communication. These limitations suggest valuable directions for future work while not diminishing the significance of our core findings.

References

H. Samy Alim, John R. Rickford, and Arnetha F. Ball. 2016. *Raciolinguistics: How Language Shapes Our Ideas About Race*. Oxford University Press.

Michael Argyle and Luo Lu. 1990. The happiness of extraverts. *Personality and individual differences*, 11(10):1011–1017.

April Baker-Bell. 2020. *Linguistic justice: Black language, literacy, identity, and pedagogy*. Routledge.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. *arXiv preprint arXiv:2005.14050*.

Su Lin Blodgett and Brendan T. O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *ArXiv*.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*.

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.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. In *NeurIPS*.

Wendi Cui, Jiaxin Zhang, Zhuohang Li, Lopez Damien, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2024. Dcr-consistency: Divide-conquer-reasoning for consistency evaluation and improvement of large language models. *arXiv preprint arXiv:2401.02132*.

N. Deas, J. Grieser, S. Kleiner, D. Patton, E. Turcan, and K. McKeown. 2023. Evaluation of african american language bias in natural language generation. In *Proceedings of EMNLP*.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

ME Francis and Roger J Booth. 1993. Linguistic inquiry and word count. *Southern Methodist University: Dallas, TX, USA*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and Abhishek Kadian et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Lisa J Green. 2002. *African American English: a linguistic introduction*. Cambridge University Press.

Jessica A Grieser. 2022. *The Black side of the river: Race, language, and belonging in Washington, DC*. Georgetown University Press.

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.

Abhay Gupta, Ece Yurtseven, Philip Meng, and Kevin Zhu. 2024. AAVENUE: Detecting LLM biases on NLU tasks in AAVE via a novel benchmark. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 327–333, Miami, Florida, USA. Association for Computational Linguistics.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.

Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*, 1.

Martin N Hebart and Guido Hesselmann. 2012. What visual information is processed in the human dorsal stream? *Journal of Neuroscience*, 32(24):8107–8109.

Paula Helm, Gábor Bella, Gertraud Koch, and Fausto Giunchiglia. 2024. Diversity and language technology: how language modeling bias causes epistemic injustice. *Ethics and Information Technology*, 26(1):8.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

V. Hofmann, P. R. Kalluri, D. Jurafsky, and S. King. 2024. Dialect prejudice predicts ai decisions about people's character, employability, and criminality. *arXiv preprint*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and Devendra Singh Chaplot et al. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q Jiang et al. 2023b. Generating with confidence: Uncertainty quantification for black-box large language models.

Wei Jin et al. 2024. The impact of large language models on learning.

Taylor Jones, Jessica Rose Kalbfeld, Ryan Hancock, and Robin Clark. 2019. Testifying while black: An experimental study of court reporter accuracy in transcription of african american english. *Language*, 95(2):e216–e252.

Anjali Kantharuban, Jeremiah Milbauer, Emma Strubell, and Graham Neubig. 2024. Stereotype or personalization? user identity biases chatbot recommendations. *arXiv preprint arXiv:2410.05613*.

Eliza Kosoy et al. 2023. Understanding the role of large language models in education.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2024. A survey on fairness in large language models. *Preprint*, arXiv:2308.10149.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Jing Yao, Si-Qing Chen, Michael Wooldridge, and Furu Wei. 2024. One language, many gaps: Evaluating dialect fairness and robustness of large language models in reasoning tasks. *arXiv preprint arXiv:2410.11005*.

Haoyan Luo and Lucia Specia. 2024. From understanding to utilization: A survey on explainability for large language models. *Preprint*, arXiv:2401.12874.

Kyle Mahowald et al. 2024. Dissociating language model knowledge and capabilities through instances of minimal pairs.

Eric Mitchell et al. 2023. Debate as a diagnostic tool for understanding large language model capabilities.

P. Mondorf and B. Plank. 2024. Beyond accuracy: Evaluating the reasoning behavior of large language models - a survey. *COLM 2024 Conference Proceedings*.

Baraka Muvuka, Rebekah M. Combs, Sonja D. Ayangeakaa, Naglaa M. Ali, Monica L. Wendel, and Therese Jackson. 2020. Health literacy in african-american communities: Barriers and strategies. *Health Literacy Research and Practice*, 4(3):e138–e143.

Stanislaw Jerzy Niepostyn and Wiktor Bohdan Daszczuk. 2023. Entropy as a measure of consistency in software architecture. *Entropy*, 25(2):328.

OpenAI. 2024. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Luigi Pastore and Sara Dellantonio. 2016. Modelling scientific un/certainty. why argumentation strategies trump linguistic markers use. In *Model-Based Reasoning in Science and Technology: Logical, Epistemological, and Cognitive Issues*, pages 137–164. Springer.

Richard D Rieke and Randall K Stutman. 2022. *Communication in legal advocacy*. Univ of South Carolina Press.

M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of ACL*.

Guijin Son, Sangwon Baek, Sangdae Nam, Ilgyun Jeong, and Seungone Kim. 2024. Multi-task inference: Can large language models follow multiple instructions at once? *Preprint*, arXiv:2402.11597.

Arthur Spears. 1998. African-american language use: Ideology and so-called obscenity. pages 226–250.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Ian Stewart. 2014. Now we stronger than ever: African-American English syntax in Twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 31–37, Gothenburg, Sweden. Association for Computational Linguistics.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, and Surya Bhupatiraju et al. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

D. Umerenkov, G. Zubkova, and A. Nesterov. 2023. Deciphering diagnoses: How large language models explanations influence clinical decision making. *Preprint*, arXiv:2310.01708.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. 2024. Dynamic self-consistency: Leveraging reasoning paths for efficient llm sampling. *Preprint*, arXiv:2408.17017.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *NeurIPS*.

Zhen Wu, Ritam Dutt, and Carolyn Rose. 2024. Evaluating large language models on social signal sensitivity: An appraisal theory approach. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 67–80, TBD. ACL.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, and Bowen Yu et al. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Taha Yasseri, András Kornai, and János Kertész. 2012. A practical approach to language complexity: a wikipedia case study. *PloS one*, 7(11):e48386.

Yuxuan Ye, Edwin Simpson, and Raul Santos Rodriguez. 2024. Using similarity to evaluate factual consistency in summaries. *arXiv preprint arXiv:2409.15090*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. Cobra frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of ACL*.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3701–3720.

A Appendix

A.1 Implementation Details

Psychological Processes Experiment Implementation To analyze the psychological processes in LLM-generated explanations, we employed a text analysis tool called Linguistic Inquiry and Word Count (LIWC)(Tausczik and Pennebaker, 2010). This tool identifies and categorizes words in a given text into various linguistic and psychological categories. The frequency of words within a specific category is directly related to the intensity of that category conveyed by the text. For example, a higher frequency of words associated with positive emotions indicates that the text conveys a stronger positive emotional tone.

Considering the varying lengths of explanations generated by LLMs for AAE and SAE question prompts, we standardized the absolute word frequencies for each LIWC category by calculating the frequency per 1,000 tokens. Our primary focus was on personal pronouns and psychological categories, particularly social processes, affective processes (e.g., positive emotions), cognitive processes (e.g., certainty and tentativeness), and perceptual processes.

Consistency Estimation Experiment Implementation To Implement the consistency estimation experiment, we randomly select one question prompt from each of the 57 subjects in the MMLU benchmark and from each of the 6 categories in the BigBench benchmark, resulting in a total of 63 questions. These 63 questions were fed to the LLMs, and the process was repeated 10 times to generate 10 responses for each question prompt.

To evaluate the consistency of the answers for each question, we paired the responses and calculated the BERT score for every pair. BERT score measures semantic similarity between two texts using contextual embeddings derived from a pre-trained language model like BERT(Cui et al., 2024). Given 10 responses per question, this process resulted in $\binom{10}{2} = 45$ unique pairs of answers. Ideally, if the LLM's responses are consistent, the average BERT score across these 45 pairs would be high, reflecting strong semantic alignment. On the other hand, lower BERT scores would indicate inconsistency among the responses generated by the LLM. We selected BERT score as our metric because it assesses similarity based on contextual meaning rather than relying only on exact word matches. This makes it a more robust measure for evaluating textual consistency.

Moreover, the parsed letter-form answer from the 10 answers provide additional insight into the consistency of the LLM's ability to produce accurate responses. To evaluate this, we use entropy as a measure of the purity of the answers (Farquhar et al., 2024):

$$H = -\sum_{i=1}^{n} p_i \log_b(p_i)$$

lower entropy indicates higher consistency, while higher entropy suggests greater variability in the LLM generated answers(Niepostyn and Daszczuk, 2023).

A.2 Prompts and Engineering Details

Environments Our experiments were conducted using Python 3.11.8 as the primary programming environment. The core analysis relied on several key libraries: Transformers (4.47.0) for model implementations, Langchain (0.3.11) for large language model interactions, and Datasets (3.2.0) for efficient data handling. We utilized Scikit-learn (1.6.0) and SciPy (1.14.1) for statistical analysis, and Pandas (2.2.3) for data manipulation. For visualization, we employed Matplotlib (3.9.4) and Seaborn (0.13.2). For hardware infrastructure, we deployed open-source models on NVIDIA A100 GPUs, while GPT family models were accessed through Azure OpenAI services. Detailed dependencies and configurations are available in our public repository.

Question Prompt Generation The first step of our experiment is to generate the question prompts that simulate real world Q&A interaction between a user and LLMs. To achieve this, we utilized existing benchmarks, such as MMLU and Bigbench, which contains multiple choice questions which covers various different topics. From the MMLU benchmark, we randomly sampled 50 questions across 57 subjects, resulting in a total of 2,850 multiple-choice questions. From Bigbench benchmark, we selected 1333 multiple choice questions that are related to logical thinking such as navigation, data understanding and causal judgment, etc. We then used GPT-4.0 Turbo to generate question prompts by providing it with the original multiple-choice questions, simulating real-world users asking the LLMs these questions.

Next, we made a copy of the original 2,850 multiple-choice question prompts and converted them from Standard American English (SAE) to African American English (AAE). This setup created two groups: a control group with the original SAE prompts and an experimental group with the converted AAE prompts. Both sets of question prompts were then fed into different LLMs to generate answers and explanations.

Answer Extraction LLM-generated answers often include explanations, with the letter-form answer typically embedded within these explanations. To parse the letter-form answer for accuracy evaluation, we designed an LLM-based parser, specifically prompted to identify the option (A-D) toward the end of the explanation, which is framed with phrases like, "My final answer is: ...". For BigBench benchmark questions, which include more options, the parser was adapted to parse option from A-F.

To validate the performance of the LLM-based parser, we conducted a human evaluation. Human labeler annotated the final option output from a set of 100 LLM-generated answers, providing ground truth labels. We then let the LLM-based parser parse the same set of 100 answers. We computed Cohen's Kappa coefficient to assess the agreement between the final options parsed by LLM-based parser and the human labeler. The resulting Cohen's Kappa coefficient was 0.93, indicating nearly perfect agreement. This high score confirms the validity of our LLM-based parser.

Mitigation Strategy Prompts We detail the prompting strategies used in our experiments below. For each strategy, we provide the base prompt structure and an example. All prompts are appended to the original task instruction.

Expert Teacher. This strategy frames the model as an experienced educator who is skilled in working with diverse student populations:

> You are an experienced teacher with expertise in working with students from diverse linguistic backgrounds. Your goal is to explain concepts clearly while respecting and accommodating different language varieties. Please read the following question and provide your response: [Question Text]

Cultural Context. This approach explicitly acknowledges different linguistic and cultural contexts:

> The following question may be presented in different language varieties, including African American English (AAE) or Standard American English (SAE). Each variety has its own valid grammatical rules and cultural context. Please analyze the question with respect to its linguistic context: [Question Text]

Dialect Recognition. This strategy directly instructs the model to process different language varieties appropriately:

> When responding to this question, be aware that it may be expressed in different English dialects. Apply your understanding of dialect-specific features and grammatical patterns. Consider all dialectal variations as equally valid forms of expression: [Question Text]

	Model Response Characteristics	
Metrics	LLaMA 3.1	Uncensored LLaMA 3.1
Accuracy (%)	47.8	47.3 (-1.0%↓)
FRES Score	57.3	63.6 (+11.0%↑)

Table 6: Comparison between safeguarded and uncensored versions of LLaMA 3.1 (8B). While accuracy shows minimal decline in the uncensored version, removing safety measures leads to substantial increases in readability complexity, suggesting amplification of response patterns when safety constraints are removed. The percentages indicate relative changes from the base model.

Readability Focus. This approach emphasizes clear communication while maintaining consistent comprehension across dialects:

> Please ensure your response is clear and accessible across different English varieties. Focus on maintaining consistent meaning and comprehension regardless of the dialect used. Analyze the following question: [Question Text]

Multi-strategy. This comprehensive approach combines elements from the above strategies:

> As an experienced educator skilled in working with diverse linguistic backgrounds, please address this question while: 1. Recognizing and respecting different language varieties (including AAE and SAE) 2. Ensuring clear communication across dialects 3. Maintaining consistent comprehension 4. Acknowledging the validity of different grammatical patterns
>
> Please analyze the following question: [Question Text]

These prompting strategies were designed to systematically address potential dialectal biases while maintaining the model's ability to effectively process and respond to questions. Each strategy was applied consistently across all test cases to ensure comparable results.

A.3 Additional Analysis

Uncensored Model Exacerbates the Bias To investigate whether safety measures affect dialectal biases, we compared AAE responses between safeguarded and uncensored versions of Llama3.1

8B. As shown in Table??, while accuracy remains relatively stable (dropping by only 1%), removing safety measures significantly amplifies dialectal response patterns. The uncensored model shows consistently higher FRES scores (increasing by 11%) across all datasets. This suggests that model safeguards may actually help moderate the model's tendency to adjust its response style based on dialect, and their removal leads to more exaggerated dialectal adaptations.

Readability of the Uncensored Model The models we used in this study are all popular LLMs that are heavily safeguarded, yet we still observed a significant discrepancy in readability. Our hypothesis is that uncensored models would exhibit an even greater discrepancy in readability, which would make the bias appear more pronounced. To test this hypothesis, we employed an uncensored Llama3.1 8B model and compared its performance with the safeguarded Llama3.1 8B model on the same set of AAE question prompts. The results showed that the FRES scores of explanations generated by the uncensored Llama3.1 8B model for AAE question prompts were even higher compared to those generated by the safeguarded version. This put the explanations from the uncensored model into even lower grade-level readability categories. These findings suggest that LLMs tend to provide easier and more readable answers to questions written in AAE compared to SAE, creating a significant readability discrepancy. Furthermore, the lack of safeguarding mechanisms in LLMs appears to exacerbate this discrepancy in readability (graph to be added later).

A.4 Human Validation

To validate our findings, we recruited 15 native African American English (AAE) speakers to serve as annotators. All participants provided their consent to participate in this study by signing the consent form. Four separate surveys were designed, each containing 25 questions aimed at evaluating metrics such as fluency, coherence, understandability, and overall quality, based on the AAVENUE framework (Gupta et al., 2024), as illustrated in Fig 4. Each survey was completed by three annotators, involving a total of 12 annotators for the task. For the realism score annotation, an additional set of three annotators assessed the realism of 25 questions as illustrated in Fig 5.

We estimated that each survey would take approximately 20–25 minutes to complete. Annotators were compensated $7 for each task, equating to an hourly rate of approximately $21/hour. This ensured fair payment for their time and effort.

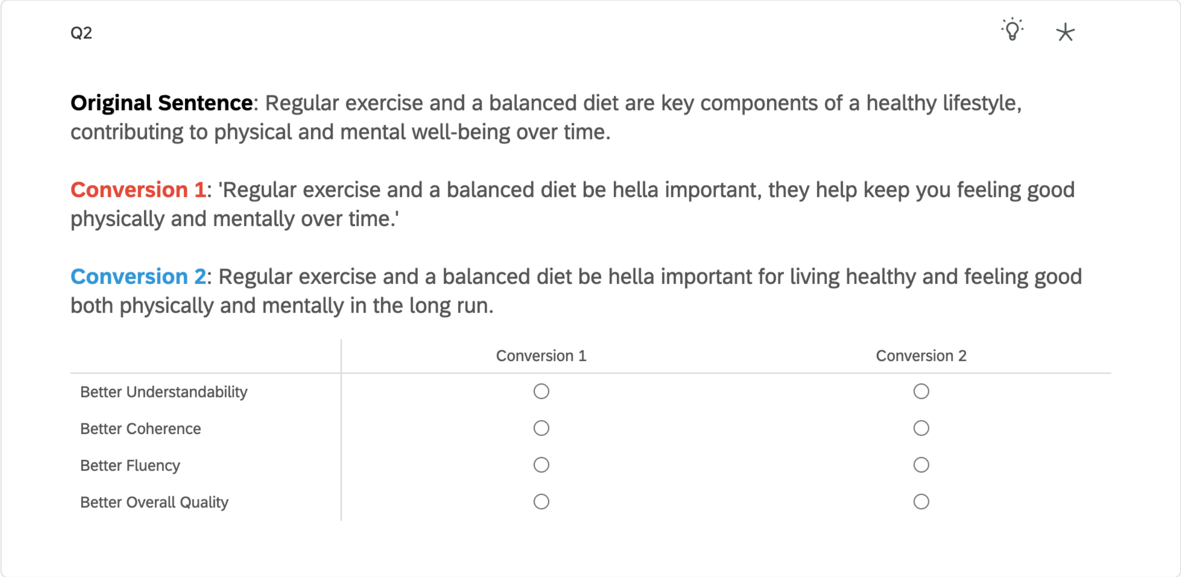
In addition, This study was approved by our Institutional Review Board (IRB) and all participants provided informed consent.

Models	Avg. Explanation Length(AAE)	Avg. Explanation Length(SAE)	T-statistic	P-value
GPT-4	352.52	352.02	0.22	0.82
GPT-3.5	170.16	171.60	-0.48	0.14
Llama3.1 8B *	362.82	358.99	1.21	0.02
Llama3.2 3B	250.00	247.21	0.37	0.57
qwen2.5 7B	262.20	260.06	0.83	0.40
gemma2 9B **	167.88	162.56	1.56	0.01
mistral 7B	229.34	230.99	-0.34	0.63

Table 7: The comparison of LLM explanation text lengths for SAE and AAE prompts shows no significant differences for most tested models.

feature	explanation	example	standard english
Auxiliaries	AAE allows copula deletion	We better than before.	We are better than before.
Completive done	To Indicate completion	I done wrote it.	I had written it.
The word "ass"	It can appear reflexively	Get yo'ass inside!	get inside!
Existential it	To indicate something exists	It's some milk in the fridge	There is some milk in the fridge.
Future gonna	To mark future tense	You gonna understand	You are going to understand
Got	Can replace the verb form of have	I got to go	I have to go
No Inflection	Certain tense don't need inflection	She study linguistics	She studies linguistics
Negative concord	NPIs agree with negation	He don't have no camera	He doesn't have a camera
Negative inversion	Similar to negative concord	don't nobody never say	nobody ever says
Null genitives	Drop any possessive endings	Rolanda bed	Rolanda's bed
Habitual be	marks habitual action	he be in his house	he is in his house

Table 8: Complete set of lexical and morphosyntactic features with examples mentioned in VALUE benchmark



Q2

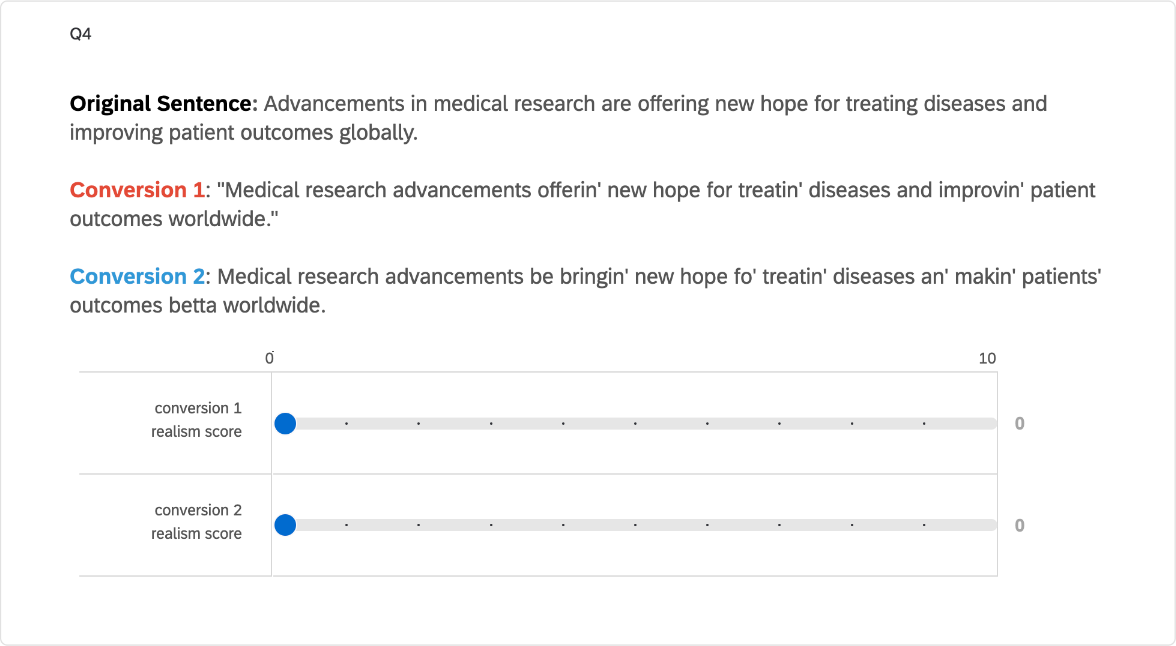
Original Sentence: Regular exercise and a balanced diet are key components of a healthy lifestyle, contributing to physical and mental well-being over time.

Conversion 1: 'Regular exercise and a balanced diet be hella important, they help keep you feeling good physically and mentally over time.'

Conversion 2: Regular exercise and a balanced diet be hella important for living healthy and feeling good both physically and mentally in the long run.

	Conversion 1	Conversion 2
Better Understandability	○	○
Better Coherence	○	○
Better Fluency	○	○
Better Overall Quality	○	○

Figure 4: Sample question that we ask annotator to rank the converted AAE and SAE sentences based on certain metrics.



Q4

Original Sentence: Advancements in medical research are offering new hope for treating diseases and improving patient outcomes globally.

Conversion 1: "Medical research advancements offerin' new hope for treatin' diseases and improvin' patient outcomes worldwide."

Conversion 2: Medical research advancements be bringin' new hope fo' treatin' diseases an' makin' patients' outcomes betta worldwide.

0 10

conversion 1 realism score 0

conversion 2 realism score 0

Figure 5: Sample question that we ask annotator to realism of the converted AAE and SAE sentences on a scale from 0-10

Metric	**Survey 1**	**Survey 2**	**Survey 3**	**Survey 4**	**Average**
Understandability (Ours)	86.6%	68.0%	65.3%	80.3%	75.1%
Understandability (SotA)	13.4%	32.0%	34.7%	19.7%	24.9%
Coherence (Ours)	84.0%	72.5%	66.6%	83.3%	76.6%
Coherence (SotA)	16.0%	27.5%	33.3%	16.7%	23.4%
Fluency (Ours)	85.3%	70.6%	75.0%	84.0%	78.7%
Fluency (SotA)	14.7%	29.4%	25.0%	16.0%	22.3%
Overall Quality (Ours)	85.3%	64.0%	69.3%	76.3%	73.7%
Overall Quality (SotA)	14.7%	36.0%	30.7%	23.7%	26.3%

Table 9: Human evaluation results comparing our LLM-based dialect conversion method to the SotA baseline (Gupta et al., 2024) across four surveys (S1-4). Each cell shows the % of evaluators who prefer that method.