Disparities in LLM Accuracy and Reasoning: A Case Study on African American English

Anonymous authors
Paper under double-blind review

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in reasoning 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). We systematically investigate dialectal disparities in LLM reasoning tasks. To do this, we compare LLM performance on Standard American English (SAE) and AAE prompts, building on the recent LLM-based dialect conversion methods and enabling linguistic analyses on the LLMs' reasoning. We find that LLMs consistently produce less-accurate responses and simpler reasoning for AAE prompts compared to their SAE equivalents on widely-used benchmarks. With expert evaluation, we also find that our prompts adequately reflect AAE, suggesting dialectal disparities may already be biasing deployed systems.[1]

1 Introduction

Large Language Models (LLMs) have increasingly demonstrated sophisticated reasoning capabilities, with techniques like chain-of-thought prompting Wei et al. (2022) enabling models to articulate their reasoning process before providing answers. This advancement has unlocked valuable applications beyond simple task completion, particularly in educational contexts Kosoy et al. (2023) where learners can examine and learn from model explanations. As these reasoning-capable models rapidly expand into high-stakes domains like healthcare Wu et al. (2025) and education Wambsganss et al. (2023), ensuring their utility doesn't depend on a user's dialect or language variety becomes increasingly critical to avoid amplifying existing educational and professional disparities.

Prior research has established that language technologies exhibit systematic biases against non-dominant language varieties, particularly African American English (AAE) Blodgett et al. (2020); Green (2002). Performance disparities among various dialects have been documented across tasks including toxicity detection Sap et al. (2019), text generation Groenwold et al. (2020), and language identification Blodgett & O'Connor (2017) for LLMs. However, these studies primarily focus on output accuracy without examining how dialectal variations impact the reasoning process, which has becoming increasingly relevant given the advancement of reasoning if LLMs Mondorf & Plank (2024b). This gap introduces unexplored risks: models might provide correct answers with flawed reasoning paths, offer qualitatively different explanation styles across dialects, or implicitly reinforce linguistic hierarchies by treating certain dialects as less deserving of rigorous reasoning.

To address the critical gap in understanding how dialectal variations affect LLM reasoning processes—beyond mere output accuracy, we develop an experimental framework that investigates three fundamental questions: (1) Do LLMs exhibit different reasoning capabilities and accuracy when processing AAE versus SAE inputs? (2) How do explanation structure, complexity, and linguistic features differ across dialects? (3) Are LLMs equally consistent and reliable in their reasoning regardless of input dialect? As illustrated in Figure 1, we answer these questions by creating paired SAE-AAE inputs using LLM-based dialect transformation, which we validate with AAE speakers who rate the converted texts as natural compared to prior approaches. We systematically evaluate 7 recent LLMs across these paired inputs, measuring performance disparities in accuracy, explanation quality, and response consistency. This approach allows us to isolate the effect of dialect while

[1] All of our code and data are publicly-available: https://anonymous.4open.science/r/dialect_bias_eval-8FCB/readme.md

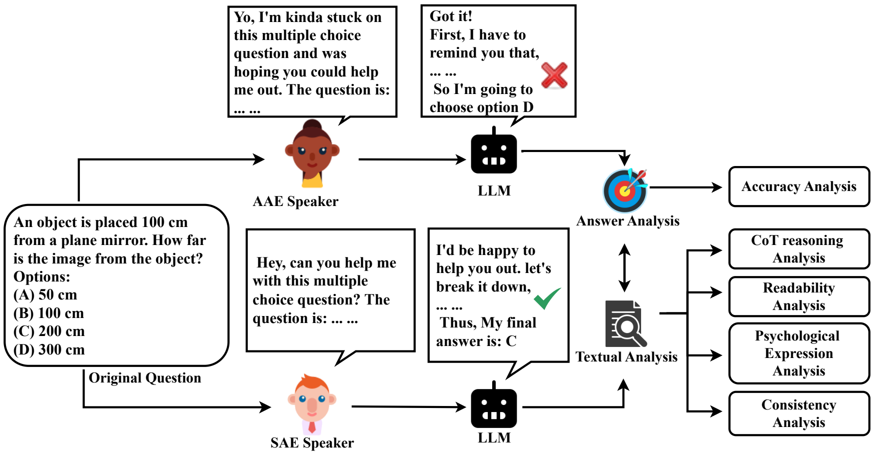


Figure 1: The experiment simulates a question-and-answer session to evaluate potential language model biases when responding to different English dialects. Specifically, it compares the accuracy and consistency of responses to prompts written in African American English (AAE) versus Standard American English (SAE). The study also analyzes the explanations provided in SAE, as it is the case in many applications, examining their consistency, readability, and psychological expression.

controlling for semantic content, providing insight into how linguistic variation affects each stage of LLM reasoning from initial interpretation to explanation generation and final answer selection.

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). In sum, our contributions are: (1) We develop a systematic framework to evaluate how LLMs reason across dialects, combining dialect conversion with reasoning analysis to assess model behavior with different language varieties. (2) We present the first comprehensive analysis of dialectal bias in LLM reasoning, revealing disparities in accuracy, explanation sophistication, readability, and linguistic patterns. (3) We identify effective prompting strategies that reduce these dialectal disparities while preserving model performance, providing practical solutions for more equitable LLM deployment.

2 Background & Related Work

2.1 Background: African American English and Language Disparity

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).[2] This hierarchical view of language varieties reflects and perpetuates broader societal biases, particularly affecting AAE speakers in contexts like education, housing, employment, and the criminal justice system Adger et al. (2014); Rickford & King (2016); Massey & Lundy (2001); Grogger (2011).[3]As language technologies increasingly serve broader populations Milmo

[2]AAE can be referred as African American Vernacular English (AAVE) or African American Language (AAL) Grieser (2022), Similarly, SAE, i.e., the dominant or canonical variant of American English, is can be referred to as White Mainstream English (WME) or Mainstream US English (MUSE) (Sap et al., 2019; Kantharuban et al., 2024).

[3]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.

(2023); La Malfa et al. (2024), addressing anti-AAE bias is essential for advancing linguistic justice and ensuring equitable access to AI systems Li et al. (2024); Alim et al. (2016).

2.2 Dialect Bias in NLP Systems

NLP systems exhibit systematic biases against non-standard 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 limitations in scalability and analysis depth. Recent work by Lin et al. (2024) demonstrates LLMs' brittleness to dialects in reasoning tasks, persisting across architectures and prompting techniques Kojima et al. (2022); Wei et al. (2022), while Li et al. (2025) investigates implicit biases through agent-based simulations. Our work advances this research by: (1) developing automated methods for dialect-aware evaluations, (2) conducting more comprehensive evaluation with diverse metrics beyond accuracy measures Mondorf & Plank (2024a); Wan et al. (2024), (3) analyzing conversational norms affecting model performance, and (4) proposing mitigation strategies based on fine-grained analysis of reasoning processes across dialects 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 built 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 Comparison with Existing 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 tends 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 develop a more systematic and linguistically-grounded conversion method by using a rigorously structured prompt (provided in Appendix A.2) that systematically incorporates 11 key morphosyntactic features described in the VALUE benchmark. Unlike previous methods that rely on arbitrary few-shot demonstrations, our prompt provides explicit translation rules with linguistically-grounded examples for each feature, ensuring consistent and principled conversion. These rules and examples are 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 Appendix A.2 and Table 9. This approach is designed to improve the converter's performance by offering a more comprehensive and linguistically representative corpus of AAE text patterns.

3.2 Human evaluation

To validate our dialect conversion approach, we conduct a human evaluation using 100 SAE sentences generated by GPT-4. We convert these sentences into AAE using two methods: a state-of-the-art (SotA) LLM-based dialect converter introduced by AAVENUE benchmarkGupta et al. (2024) and our own LLM-based converter. We then recruit native AAE speakers from Prolific to rank the AAE conversions from each method in terms of **fluency**, **coherence**, **understandability**, and **overall quality**. **Fluency** assesses the grammatical correctness and writing quality of the generated text;

Coherence evaluates the logical flow and consistency of ideas within the translations; **Understandability** measures how easily readers could comprehend the translation, and **Quality** offers a holistic evaluation of the overall standard of the text. We also use Fleiss' κ to assess inter-annotator agreement across the four metrics, we find that annotators agreed substantially Kılıç (2015). Additional details and ethical consideration are mentioned in the Appendix A.3.

The result from Figure 3 in the appendix shows that our dialect conversion method significantly outperformed the SotA AAVENUE converter, achieving a substantial margin of preference across all evaluated metrics (74–79% win rate over AAVENUE). Statistical significance is assessed via paired binomial tests on aggregated pairwise preferences of 25 converted AAE sentences for all annotators, with complete results shown in Table 10. Additionally, we recruit native AAE speakers to rate converted AAE sentences for realism (0-10 scale), where our method scores 7.97/10 ($\pm$0.21) compared to the state-of-the-art's 7.62/10 ($\pm$0.28). With an inter-annotator agreement of 0.61, these results validate our approach's effectiveness in producing realistic AAE translations.

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 answers generated by the LLMs for both SAE and AAE questions were validated as Standard American English (SAE) using external tools.[4]. All of the implementation details of the following metrics can be found in Appendix A.1 and Appendix A.2.

4.1 Evaluation Metrics

Accuracy The most direct measurement of LLM answer quality is the answer accuracy. To calculate this, we use an LLM-based parser to parse the letter-form answer from the generated explanations as shown in Figure 1. We then calculate 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 a higher readability for one dialect could signal oversimplification at the expense of depth or complexity Yasseri et al. (2012).

To assess readability, we employ 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 use the Linguistic Inquiry and Word Count (LIWC) tool Tausczik & Pennebaker (2010); Francis & 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 8, we still standardize linguistic marker frequencies to per 1,000 words. This ensures a co-comparable analysis of linguistic features across the two dialects.

[4]We used a model trained on Twitter data to classify whether each response was in AAVE or SAE, which follows the approach of Blodgett et al. (2016)Blodgett et al. (2016). The results show that 99.92% of the generated answers were classified as SAE.

	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. A paired T-test is performed on each model to assess statistical significance. Statistically significant results are marked with ** ($p < 0.01$) and * ($0.01 <= p < 0.05$). The full statistical test results are presented in Table 7.

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 estimate consistency, we randomly sample 100 multiple choice question prompts based on the MMLU dataset and generate 10 outputs for each sampled question prompts and measured variability in their content and quality. If the LLMs 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. (2023). 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 groupe 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 Appendix 3.To ensure fair evaluation, a reversion test (AAE to SAE, see A.3) demonstrated minimal information loss (93.7% semantic equivalence).

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, represents 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's explainability. Luo & Specia (2024).

5 Main Results

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

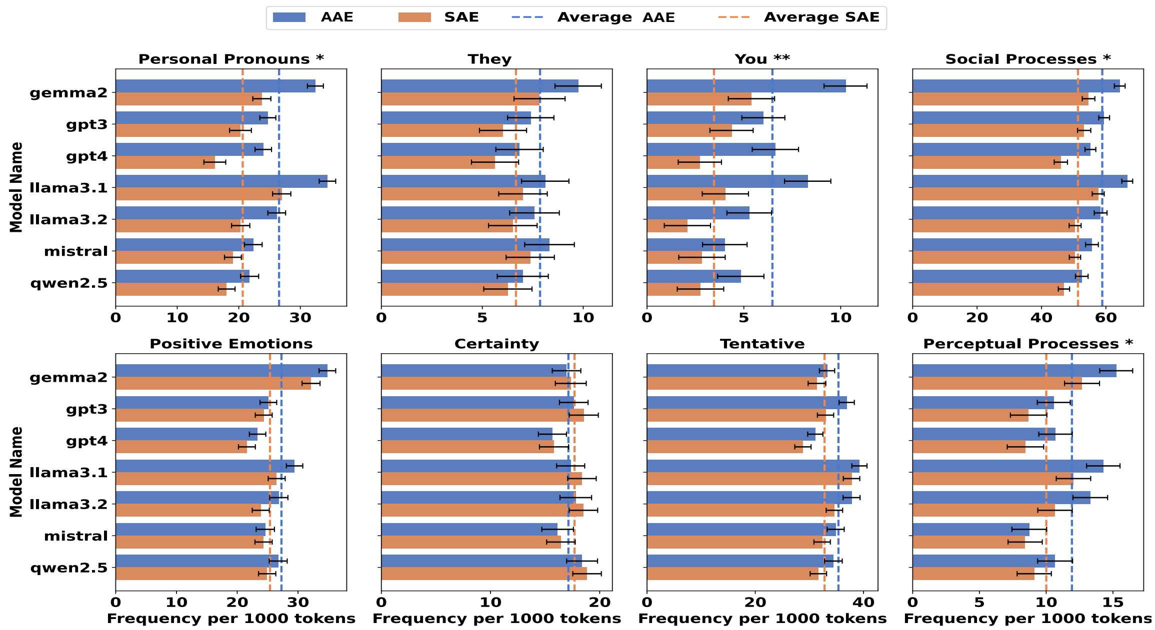


Figure 2: Linguistic Marker Differences in Explanations for AAE and SAE Prompts: Frequencies of linguistic markers, calculated by LIWC and standardized per 1 K tokens; marked with ** and * for statistical significance (**: $p < 0.01$, *: $0.01 <= p < 0.05$).

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 prompts? To assess readability, we utilize the Flesch Reading Ease Score (FRES) Flesch (1948). The detailed FRES shown in Table 5 indicate a statistically significant difference in the complexity of language used in explanations generated by LLMs for SAE and AAE prompts. Specifically, explanations for SAE prompts tend to correspond to college-level readability (FRES below 50), whereas those for AAE prompts more often align with a 12th-grade level or lower (FRES of 50 or higher). This discrepancy suggests that LLMs generate more complex, formal, and academically structured responses for SAE inputs, while AAE responses may be comparatively simplified or less sophisticated. Such a pattern may indicate underlying biases in the training data or language modeling process Deas et al. (2023).

RQ3: How do the psychological expressions in LLM-generated explanations differ between SAE and AAE question prompts? Our analysis (Figure 2) 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

	Entropy (↓)		BERT Score (↑)		Average Acc. (↑)	
Models	**SAE**	**AAE**	**SAE**	**AAE**	**SAE**	**AAE**
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)**
Llama 3.1 8B	0.70	1.10 **(+0.40)**	0.85	0.81 **(-0.04)**	0.72	0.58 **(-0.14)**
Llama 3.2 3B	0.97	1.24 **(+0.27)**	0.85	0.82 **(-0.03)**	0.61	0.49 **(-0.12)**
Qwen 2.5 7B	0.41	0.84 **(+0.43)**	0.87	0.85 **(-0.02)**	0.79	0.66 **(-0.13)**
Gemma 2 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 2: Comparison of output consistency across SAE and AAE question prompts for various LLMs using (1) entropy of answers , (2) BERT Score between answer pairs, and (3) average accuracy. 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 3: 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 **bold red** indicating larger differences in rationalization.

explanations may indicate an emphasis on social connection and relational communication Argyle & Lu (1990). The greater use of perceptual process words also suggests that AAE explanations might favor more concrete reasoning Rieke & Stutman (2022); Pastore & 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 contextsHebart & 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 processesHelm 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 2) 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 & 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.

Strategy	Accuracy (%)		FRES Score	
	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 4: Comparison of prompting strategies for mitigating dialectal biases in GPT-3.5 and GPT-4. **SAE**: Standard American English; **AAE**: African American English. FRES scores range from 0–100, with higher values indicating greater readability. Differences between AAE and SAE results are shown in parentheses, with improvements in green and decreases in red.

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 3. 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 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 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 analyses in Section 5 demonstrate significant performance and explanation disparities between SAE and AAE inputs. To address these findings—lower accuracy (RQ1), simpler explanations (RQ2), different linguistic patterns (RQ3), less consistency (RQ4), and bias amplification in post-hoc rationalization (RQ5)—we test several prompting strategies designed to mitigate these specific biases.

We implement four distinct approaches, each targeting different aspects of the observed disparities: **Expert Teacher Framing** presents the LLM with an educator persona skilled in working with diverse linguistic backgrounds. Inspired by work from Zheng et al. (2024) which suggests persona in prompts can be helpful to elicit relevant knowledge, we apply this strategy because expert framing reduces linguistic bias, addressing the simpler explanations and different linguistic patterns observed with AAE. **Cultural Contextualization** explicitly acknowledges AAE as a valid language variety with its own grammatical rules, building on Sap et al. (2019)'s work showing that cultural context reduces racial bias in explanation. This is implemented by instructing the model to analyze questions with respect to their linguistic context, thus to improve its overall performance and reduce bias. **Dialect Recognition** directly instructs the model to process different dialects appropriately, targeting consistency issues by having the model apply understanding of dialect-specific features. **Readability Focus** instructs the model to maintain consistent explanation complexity across dialects, directly addressing the readability disparity. We also test a **Multi-strategy Approach** that combines elements from all four strategies into a comprehensive prompt. More details and exact prompts for each strategy are provided in Appendix A.2.

Table 4 shows these strategies are generally effective. Expert Teacher framing improves performance for both dialects but yields larger gains for AAE, reducing the performance gap from 10.7% to 7.9%. This suggests that positioning the model as an expert helps overcome biases about appropriate complexity levels for different dialects. Cultural contextualization and dialect recognition show a trade-off: slightly decreasing SAE performance while improving AAE performance, reducing the overall gap. This implies that explicitly acknowledging dialectal differences improves equity, even with a small cost to majority dialect performance. Readability-focused prompting primarily affects readability metrics, reducing the FRES score gap by nearly half. The multi-strategy approach yields the most comprehensive improvements in both accuracy and linguistic markers. However, Son et al. (2024) show that LLMs often struggle with multiple simultaneous instructions, potentially limiting the scalability of combined interventions and can be explored further as future research direction.

7 Conclusion and Limitations

This work systematically investigates dialectal disparities in LLM reasoning, revealing significant variations in the processing of AAE and SAE inputs. Our findings demonstrate a fundamental influence of dialectal bias on the construction of logical arguments, affecting performance metrics, reasoning sophistication, and the potential for stereotype expression. While advancements in model scaling and training have yielded improvements in general reasoning capabilities, persistent disparities across dialects suggest that a more nuanced approach to fairness is required. We argue for the essential consideration of dialectal fairness in LLM development and training, particularly in reasoning-intensive applications where such biases may remain latent and thus carry substantial implications.

Our study has important limitations to consider. Conceptually, we focus 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.

Additionally, we acknowledge limitations in our comparison of dialect conversion methods. Our evaluation is based on a sample of 100 converted sentences, annotated by 12 raters—a relatively small sample size that limits statistical power. Future work will expand this evaluation to provide a more robust assessment of our dialect conversion approach.

Ethical Considerations

This study was conducted with careful attention to ethical considerations in research involving linguistic minorities. All human evaluation procedures were approved by our Institutional Review Board (IRB), and we obtained informed consent from all participants. To respect and accurately represent AAE as a legitimate, rule-governed language variety, we consulted linguistic research and engaged native AAE speakers in validating our dialect conversion methodology. We acknowledge that research involving minority language varieties requires particular sensitivity to avoid perpetuating linguistic discrimination. Our findings about performance disparities between AAE and SAE are presented with the explicit goal of identifying and addressing systematic biases in language models, rather than suggesting inherent advantages of one language variety over another. The code and data from this study will be made publicly available to ensure reproducibility and facilitate further research on making language models more equitable across dialects.

References

Carolyn Temple Adger, Walt Wolfram, and Donna Christian. *Dialects in schools and communities*. Routledge, 2014.

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

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

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

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

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of EMNLP*, pp. 1119–1130, 2016.

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

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

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

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

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

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

Rudolf Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948. doi: 10.1037/h0057532. URL https://doi.org/10.1037/h0057532.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and Abhishek Kadian et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

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

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

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. Investigating african-american vernacular english in transformer-based text generation. *Proceedings of EMNLP*, 2020.

Jeffrey Grogger. Speech patterns and racial wage inequality. *Journal of Human Resources*, 46(1): 1–25, 2011.

Abhay Gupta, Ece Yurtseven, Philip Meng, and Kevin Zhu. AAVENUE: Detecting LLM biases on NLU tasks in AAVE via a novel benchmark. In Daryna Dementieva, Oana Ignat, Zhijing Jin, Rada Mihalcea, Giorgio Piatti, Joel Tetreault, Steven Wilson, and Jieyu Zhao (eds.), *Proceedings of the Third Workshop on NLP for Positive Impact*, pp. 327–333, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4pi-1.28. URL https://aclanthology.org/2024.nlp4pi-1.28.

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

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

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

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

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

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and Devendra Singh Chaplot et al. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

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

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

Selim Kılıç. Kappa testi. *Journal of mood disorders*, 5(3):142–144, 2015.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

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

Emanuele La Malfa et al. Language-models-as-a-service: Overview of a new paradigm and its challenges. *Journal of Artificial Intelligence Research*, 80:1497–1523, 2024.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models, 2024. URL https://arxiv.org/abs/2308.10149.

Yuxuan Li, Hirokazu Shirado, and Sauvik Das. Actions speak louder than words: Agent decisions reveal implicit biases in language models, 2025. URL https://arxiv.org/abs/2501.17420.

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

Haoyan Luo and Lucia Specia. From understanding to utilization: A survey on explainability for large language models, 2024. URL https://arxiv.org/abs/2401.12874.

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

Douglas S Massey and Garvey Lundy. Use of black english and racial discrimination in urban housing markets. *Urban Affairs Review*, 36(4):452–469, 2001.

Dan Milmo. Chatgpt passes 100 million users, making it the fastest-growing app in history. *The Guardian*, 2023.

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

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

Philipp Mondorf and Barbara Plank. Beyond accuracy: Evaluating the reasoning behavior of large language models - a survey. In *First Conference on Language Modeling*, 2024b. URL https://openreview.net/forum?id=Lmjgl2n11u.

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

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

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

John R Rickford and Sharese King. Language and linguistics on trial: Hearing rachel jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, 92(4):948–988, 2016.

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

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

Guijin Son, Sangwon Baek, Sangdae Nam, Ilgyun Jeong, and Seungone Kim. Multi-task inference: Can large language models follow multiple instructions at once?, 2024. URL https://arxiv.org/abs/2402.11597.

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

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. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

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

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, and Surya Bhupatiraju et al. Gemma: Open models based on gemini research and technology, 2024. URL https://arxiv.org/abs/2403.08295.

Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Neshaei, Roman Rietsche, and Tanja Käser. Unraveling downstream gender bias from large language models: A study on AI educational writing assistance. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10275–10288, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.689. URL https://aclanthology.org/2023.findings-emnlp.689/.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. Dynamic self-consistency: Leveraging reasoning paths for efficient llm sampling, 2024. URL https://arxiv.org/abs/2408.17017.

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

Yuqi Wu, Guangya Wan, Jingjing Li, Shengming Zhao, Lingfeng Ma, Tianyi Ye, Ion Pop, Yanbo Zhang, and Jie Chen. Proai: Proactive multi-agent conversational ai with structured knowledge base for psychiatric diagnosis, 2025. URL https://arxiv.org/abs/2502.20689.

Zhen Wu, Ritam Dutt, and Carolyn Rose. Evaluating large language models on social signal sensitivity: An appraisal theory approach. In Nikita Soni, Lucie Flek, Ashish Sharma, Diyi Yang, Sara Hooker, and H. Andrew Schwartz (eds.), *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pp. 67–80, TBD, August 2024. ACL. doi: 10.18653/v1/2024.hucllm-1.6. URL https://aclanthology.org/2024.hucllm-1.6.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, and Bowen Yu et al. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

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

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

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. A survey of large language models, 2024. URL https://arxiv.org/abs/2303.18223.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15126–15154, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.888. URL https://aclanthology.org/2024.findings-emnlp.888.

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

A Appendix

A.1 Implementation Details

Psychological Processes Experiment Implementation To analyze the psychological processes in LLM-generated explanations, we employ a text analysis tool called Linguistic Inquiry and Word Count (LIWC)Tausczik & 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 standardize 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 are fed to the LLMs, and the process is repeated 10 times to generate 10 responses for each question prompt.

To evaluate the consistency of the answers for each question, we pair the responses and calculate 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 BERTCui et al. (2024). Given 10 responses per question, this process results 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 select 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 answersNiepostyn & Daszczuk (2023).

A.2 Prompts and Engineering Details

Dialect Conversion Prompt Our dialect conversion system uses the following structured prompt for consistent and linguistically-informed translation:

> Please translate the following sentence: '{sentence}' using the 13 translation rules provided as references:
>
> 1. Auxiliaries: AAE allows copula deletion (e.g.: We are better than before → We better than before.)
> 2. Completive done: this indicates completion (e.g.: I had written it. → I done wrote it.)
> 3. The word "ass": It can appear reflexively (e.g.: get inside! → Get yo'ass inside!)
> 4. Existential it: to indicate something exists (e.g: There is some milk in the fridge. → It's some milk in the fridge)

> 5. Future gonna: to mark future tense (e.g.: You are going to understand → You gonna understand)
> 6. Got: can replace the verb form of have (e.g.: I have to go → I got to go)
> 7. No Inflection: Certain tense don't need inflection (e.g.: She studies linguistics → She study linguistics)
> 8. Negative concord: NPIs agree with negation (e.g.: He doesn't have a camera → He don't have no camera)
> 9. Negative inversion: Similar to negative concord (e.g: nobody ever says → don't nobody never say)
> 10. Null genitives: Drop any possessive endings (e.g.: Rolanda's bed → Rolanda bed)
> 11. Habitual be: marks habitual action (e.g.: he is in his house → he be in his house)
>
> Your output must follow these guidelines:
> 1. Only provide the translation. Do not mention or explain how the translation was done.
> 2. Do not mention any of the 13 rules in your translation.
> 3. Format the output exactly like this: 'The translation is: ...'
> 4. Ensure the sentence sounds natural and realistic in AAE.

Environments Our experiments are conducted using Python 3.11.8 as the primary programming environment. The core analysis rely 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 utilize Scikit-learn (1.6.0) and SciPy (1.14.1) for statistical analysis, and Pandas (2.2.3) for data manipulation. For visualization, we employ Matplotlib (3.9.4) and Seaborn (0.13.2). For hardware infrastructure, we deploy open-source models on NVIDIA A100 GPUs, while GPT family models are 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 utilize existing benchmarks, such as MMLU and Bigbench, which contains multiple choice questions which covers various different topics. From the MMLU benchmark, we randomly sample 50 questions across 57 subjects, resulting in a total of 2,850 multiple-choice questions. From Bigbench benchmark, we select 1333 multiple choice questions that are related to logical thinking such as navigation, data understanding and causal judgment, etc. We then use 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 make a copy of the original 2,850 multiple-choice question prompts and convert them from Standard American English (SAE) to African American English (AAE). This setup creates two groups: a control group with the original SAE prompts and an experimental group with the converted AAE prompts. Both sets of question prompts are then fed into different LLMs to generate answers and explanations.

Answer Generation Once the LLMs receive the question prompts in different dialects, the next step is to generate responses. They're instructed to first give a detailed explanation in Standard American English (SAE) regardless of the dialect used in the question prompts, followed by the final answer choice (A, B, C, or D). Here's the prompt we used:

> "Someone asked you a multiple choice question: {question}, Please first provide an detailed explanation in Standard American English (SAE) and then choice your final answer. You need to make your explanation sounds as natural and realistic as possible. At the end, you should clearly state the letter option (A, B, C, or D) you choose. You explanation should strictly be less than 400 words.

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 design an LLM-based parser, specifically prompted to identify the option (A-D) toward

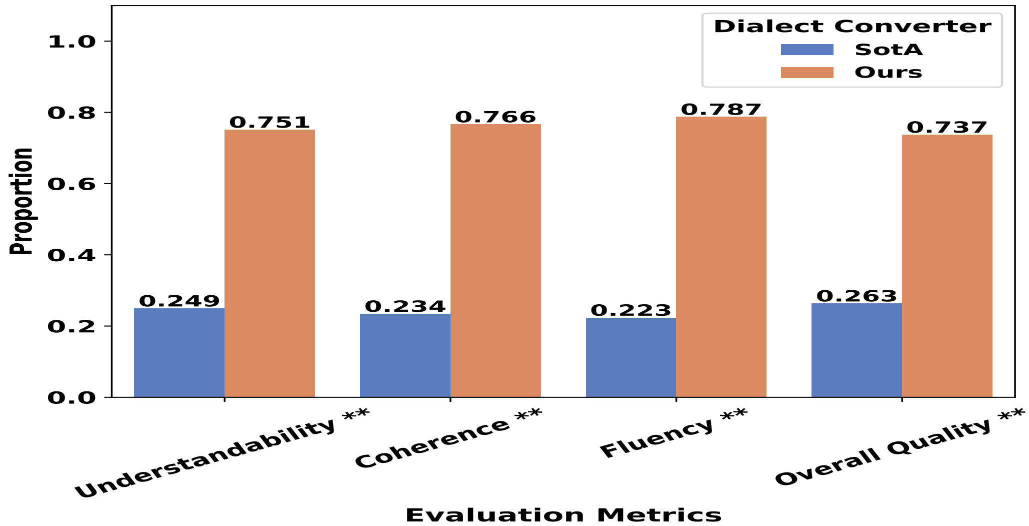


Figure 3: Average proportion of annotators favoring each SAE-AAE converter across four metrics (Gupta et al., 2024) each metric is marked with * for statistical significance (**p < 0.01).

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 is adapted to parse option from A-F.

To validate the performance of the LLM-based parser, we conduct 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 compute 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 is 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]

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 8B**	46.9±0.5	58.0±0.5	+11.1
Llama 3.2 3B**	43.8±0.6	52.4±0.5	+8.6
Qwen 2.5 7B**	45.2±0.6	50.1±0.5	+4.9
Gemma 2 9B**	51.6±0.5	62.8±0.5	+11.2
Mistral 7B**	38.7±0.6	43.3±0.6	+4.6

Table 5: Comparison of readability (FRES) in LLM responses to SAE and AAE prompts. Higher FRES indicates simpler explanations. Models marked * are statistically significant (**p < 0.01) via t-test. Normality of FRES is verified using the Shapiro-Wilk test.

	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 are designed to systematically address potential dialectal biases while maintaining the model's ability to effectively process and respond to questions. Each strategy is 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 compare AAE responses between safeguarded and uncensored versions of Llama3.1 8B. As shown in Table 6, 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 use in this study are all popular LLMs that are heavily safeguarded, yet we still observe 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 employ 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 shows that the FRES scores of explanations generated by the uncensored Llama3.1 8B model for AAE question prompts are 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.

Evaluating Information Loss in SAE-to-AAE Dialect Conversion We examine whether converting Standard American English (SAE) to African American English (AAE) results in information loss that could affect model performance. To evaluate this, we evenly sample 100 questions converted from SAE to AAE from MMLU dataset using seven LLMs in our study and then revert each of them back to SAE questions using the GPT-4.0 turbo. We analyze whether the reconverted SAE questions preserve their original meaning.

The GPT-4.0 turbo uses the following structured prompt for reverting the AAE question back to SAE question:

> Translate the following sentence from African American Vernacular English (AAVE) to Standard American English. Ensure the translation maintains the structure of the original sentence without adding extra information. The sentence is: {sentence} Format the translated sentence exactly like this: 'The translation is: ...'

Two human annotators and GPT-4.0 Turbo independently evaluate the semantic similarity between the original and reconvert SAE questions using a binary notation system (0 indicating no semantic difference, 1 indicating semantic difference). On average, 93.7% of the question pairs exhibit no semantic differences, suggesting that the SAE-to-AAE conversion introduces minimal information loss.

The average Cohen's kappa agreement rate among the three annotators (including GPT-4.0 Turbo) is 0.58, indicating moderate to substantial agreement.

To illustrate how the conversion process can occasionally introduce shifts in meaning, we present the following example, even though only a small proportion of converted questions exhibit such shifts.

Original SAE question:
"Which of the following statements about a remote procedure call is true?"
(A) It is used to call procedures with addresses that are farther than 2^{16} bytes away.
(B) It cannot return a value.
(C) It cannot pass parameters by reference.
(D) It cannot call procedures implemented in a different language.

Converted AAE question:
"Which one 'bout remote procedure call hold weight?"
(A) It call procedures where the address way over 2^{16} bytes out.
(B) It can never return no value.
(C) It won't pass parameters by reference.
(D) It can't holler at procedures made in a foreign language.

Reverted SAE question:
"Which one about remote procedure call is significant?"
(A) It calls procedures where the address is way over 2^{16} bytes out.
(B) It can never return a value.
(C) It won't be able to pass parameters by reference.
(D) It cannot invoke procedures written in a foreign language.

Human Validation To validate our findings, we recruit 15 native African American English (AAE) speakers to serve as annotators. All participants have provided their consent to participate in this

Models	T-statistic	P-value
GPT-4 *	3.259	0.047
GPT-3.5 *	4.010	0.028
Llama3.1 8B *	3.776	0.033
Llama3.2 3B	3.012	0.057
qwen2.5 7B **	8.781	0.003
gemma2 9B *	5.628	0.011
mistral 7B **	7.246	0.005

Table 7: The models that show a statistically significant difference in accuracy between AAE-prompted and SAE-prompted answers generated by LLMs are identified. A paired T-test is used to determine statistical significance. The Shapiro-Wilk test is used to check normality of the data and ensures there are no significant outliers. Results that are statistically significant are indicated with ** ($p < 0.01$) and * ($0.01 <= p < 0.05$).

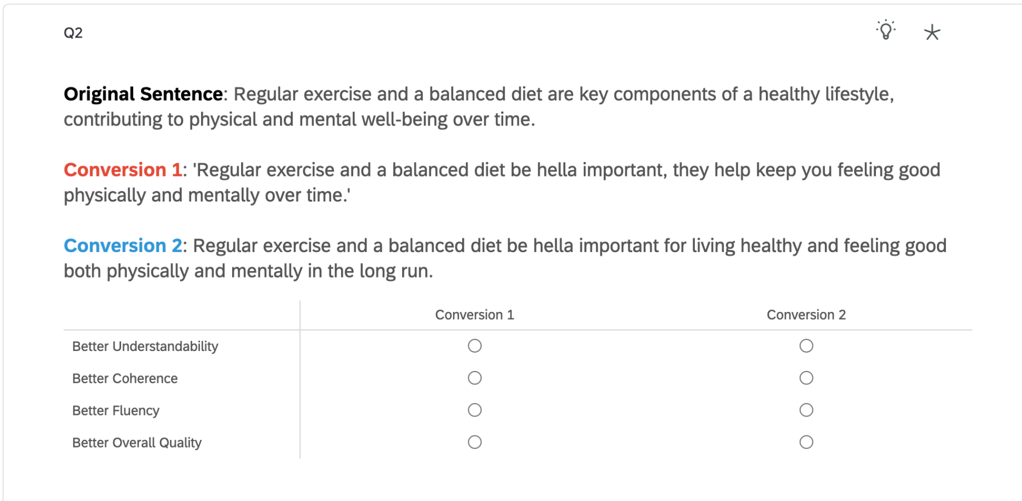
Models	Avg. Explanation Length(AAE)	Avg. Explanation Length(SAE)	T-statistic	P-value
GPT-4	352.52	352.02	0.22	0.82
GPT-3.5	170.16	171.60	-0.48	0.14
Llama3.1 8B *	362.82	358.99	1.21	0.02
Llama3.2 3B	250.00	247.21	0.37	0.57
qwen2.5 7B	262.20	260.06	0.83	0.40
gemma2 9B **	167.88	162.56	1.56	0.01
mistral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

Table 8: The comparison of LLM explanation text lengths for SAE and AAE prompts shows no significant differences for most tested models.

study by signing the consent form. Four separate surveys are 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 is 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. For the 12 annotators evaluating our dialect conversion method against state-of-the-art approaches on metrics like fluency, coherence, understandability, and overall quality, the Fleiss' κ scores are as follows: 0.65 for understandability, 0.58 for coherence, 0.62 for fluency, and 0.68 for overall quality. Additionally, for the 3 annotators assessing the realism of sentences generated by our dialect converter, the Fleiss' κ score is 0.73. This indicates moderate to substantial agreement across all evaluations.

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

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



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.

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

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 10: 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.

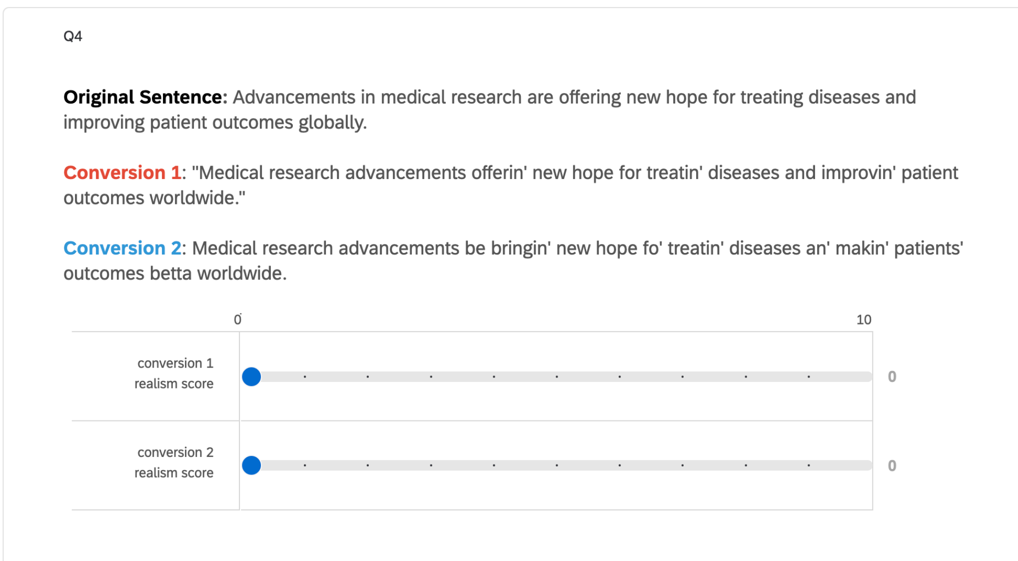


Figure 5: Sample question that we ask annotator to realism of the converted AAE and SAE sentences on a scale from 0-10