That Ain't Right: Assessing LLM Performance on QA in African American and West African English Dialects

William Coggins, Jasmine McKenzie, Sangpil Youm,
Pradham Mummaleti, Juan E. Gilbert, Eric Ragan, Bonnie J. Dorr
University of Florida, Gainesville, Florida
{william.coggins, jasminemckenzie, youms,
pradhammummaleti, juan, eragan, bonniejdorr}@ufl.edu

Abstract

As Large Language Models (LLMs) gain mainstream public usage, understanding how users interact with them becomes increasingly important. Limited variety in training data raises concerns about LLM reliability across different language inputs. To explore this, we test several LLMs using functionally equivalent prompts expressed in different English dialects. We frame this analysis using Question-Answer (QA) pairs, which allow us to detect and evaluate appropriate and anomalous model behavior. We contribute a cross-LLM testing method and a new QA dataset translated into AAVE and WAPE variants. Results show a notable drop in accuracy for one dialect relative to the baseline.

1 Introduction

Large Language Models (LLMs) are increasingly embedded in daily life, assisting users with both professional and personal tasks. Despite global use, LLMs are trained primarily on English—over 90% of which is Standard American English (SAE) (Cooper, 2023)—resulting in potential mismatches with user inputs (Dave, 2023). Popular LLMs largely train on SAE, with only about 7% of training data coming from other languages (Wiggins, 2025). Limited geographic variation can lead to misunderstanding, or hallucinations when users employ English dialects that deviate from SAE.

Consequently, LLMs do not perform equally well across speakers of different dialects. These dialects differ in vocabulary, grammar, and pronunciation, often shaped by culture. Over 30 major English dialects are spoken regularly in the U.S., and over 150 are spoken worldwide (AtlasLS, 2021).

African American Vernacular English (AAVE) and West African Pidgin English (WAPE) are two major dialects spoken by millions globally. However, they are rarely included in LLM training data. These dialects have distinct grammatical and phonological structures which increase the likelihood of LLM misinterpretation and inaccurate responses. Thus, further research is needed on how to interpret these dialects to ensure that responses are not distorted by language alternatives.

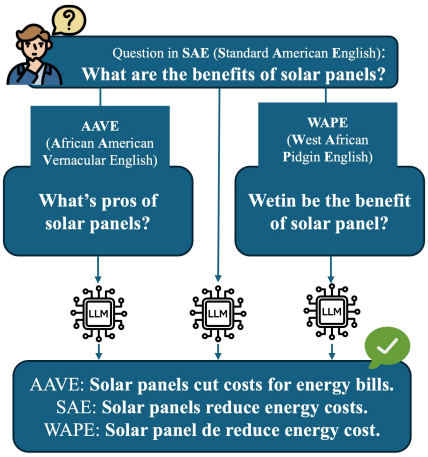

Figure 1: Dialect Translation and LLM prompting

This study focuses on *question answering* (QA) as a focused task for evaluating how well LLMs respond to prompts written in AAVE and WAPE, compared to SAE. By translating a QA dataset into these dialects and analyzing model responses, we aim to assess if LLM performance degrades with dialect input, ensuring consistent behavior across different forms of English.

By testing LLM performance on AAVE and WAPE, this study highlights where model adjustments may be needed in order to support broader consistency across a wider range of users.

2 Background and Related Work

This study examines discrepancies in LLM performance with AAVE and WAPE, building on emerging work that probes the consistency and coverage of LLMs in non-standard dialects of English. Prior

Proceedings of the 9th Widening NLP Workshop, pages 123–129

research informs our approach by demonstrating how others evaluate LLMs on different dialects, guiding how we think about achieving more consistent behavior across a broader range of inputs.

2.1 Linguistic Background

AAVE and WAPE are largely absent within LLM training data. About 30 million African Americans use AAVE in the United States (Wolfram, 2020). AAVE originates in the American South, where enslaved Africans learned English vocally. This leads to a spoken form of English that blends with Southern speech patterns. As a result, AAVE has distinctive vocabulary, grammatical structures and punctuation patterns.

WAPE, also rooted in oral traditions, is spoken by an estimated 140 million people across West African nations and in African immigrant communities originating in vocal communication (Yakpo, 2024). WAPE often features phonetic spelling, such as "them" becoming "dem," and typically omits definite and indefinite articles unless emphasis is required (Faraclas, 2017). These grammatical and lexical differences make AAVE and WAPE more likely to be misinterpreted by LLMs not exposed to them during training.

2.2 Related Work

Gupta et al. (2024) develop the AAVE Natural Language Understanding Evaluation (AAVENUE) to assess performance on natural language understanding tasks in AAVE. Our study builds on this work, examining whether performance gaps persist across dialects in a QA setting.

Lin et al. (2016) examine LLM performance on tasks like logical reasoning and math when prompts are written in AAVE. They compare model performance demonstrating noticeable drops with small differences in how the prompts are written. These findings guide our decision to directly compare the LLM outputs across dialects rather than relying on prompt translation.

Research on WAPE usage in LLMs remains limited, with newer publications and pre-prints identified. Adelani et al. (2025) develop a benchmark translating SAE to WAPE and Naija (another common Nigerian language) and test whether the WAPE-trained models also perform well for text generation. Our study takes a related approach by testing how dialects within a shared diaspora—with uneven training coverage—affect model responses. Lin et al. (2023) show that models tuned on Nigerian Pidgin outperform multilingual ones on dialect-specific tasks.

Few studies compare LLM responses to equivalent prompts across SAE, AAVE, and WAPE. Additionally, no prior work has directly compared LLM responses to the same prompts expressed in SAE, AAVE, and WAPE side by side.

3 Methodology

Dialects express similar intents in ways LLMs may interpret inconsistently. We simulate this variation by creating equivalent prompts (see Figure 1) and comparing performance against SAE, the baseline given its dominance in training data.

Assuming AAVE and WAPE yield lower QA performance than SAE, this study frames the following research questions: *To what extent do LLMs exhibit lower QA accuracy on (a) AAVE (**RQ_{AAVE}**) and (b) WAPE (**RQ_{WAPE}**), in comparison to LLM performance on equivalent SAE prompts?*

3.1 Study Design

This study aims to measure accuracy across different LLMs through a sequence of steps: selecting appropriate LLMs and a dataset to translate, applying dialect translation, and conducting evaluation (see Figure 2).

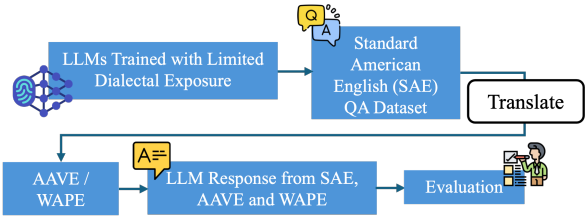

Figure 2: Methodology Flowchart

3.1.1 LLM Selection

We select LLMs based on their popularity and general accessibility to the public: ChatGPT, Grok, and Gemini. These more closely reflect current public versions of LLMs compared to enterprise models.[1]

3.1.2 Data Collection

We select the Stanford Question Dataset (SQuAD) (Rajpurkar et al., 2016) for its wide use in LLM benchmarking, objectively verifiable answers, and

[1]ChatGPT (OpenAI), Grok (xAI), and Gemini (Google) are accessed via public APIs under their respective non-commercial research terms and acceptable use policies. See Table 2 for LLM versions in Appendix A.

suitability for consistent human evaluation. Its fact-based questions allow clear accuracy judgments and help identify hallucinated outputs. With this dataset, we translate queries into AAVE and WAPE for further analysis.

3.1.3 Question Dialect Translation

We translate 1,000 randomly selected questions, using their original form as SAE prompts. AAVE versions are generated with an online dialect tool,[2] and WAPE equivalents are produced via the PidginUNMT project,[3] shown to yield high accuracy (Ogueji and Ahia, 2019). We process each SAE question using these tools to produce content-equivalent prompts.[4] A speaker of these dialects then reviews the post-translation questions.

Once validated, each question is sent to the LLM via an API, with a system statement constraining responses to fewer than 20 words This prompt is written in SAE across all dialect groups to simplify evaluation and ensure consistent comparison of short ground-truth answers with longer LLM outputs. The prompt is queried without the surrounding context featured in SQuAD since additional context creates additional ambiguity for the dialectal equivalent phrases. Each response is recorded and compared to ground truth for evaluation.

The resulting dataset consists of 1,000 SAE questions translated into AAVE and WAPE For this paper, we evaluate a random sample of 100 SAE questions and their dialectal variants, yielding 300 prompts in total. With three human raters scoring 900 responses, this represents the largest feasible scope given our resource constraints. All code used for prompting and evaluation is publicly available.[5]

3.1.4 Evaluation

We evaluate 300 prompts against ground-truth answers. These are the only ones evaluated in this study, since scoring the full 1,000 would require annotating 9,000 responses (1000 prompts $\times$ 3 dialects $\times$ 3 LLMs). The sampled questions yield 300 prompts across five domains (history, sports, religion, politics, and trivia). Each is submitted to three LLMs, producing 900 responses, which are paired with the original ground-truth answers and split across three raters for evaluation.

Raters are given 390 responses to evaluate as correct or incorrect. A 30% overlap in rater assignments balances broad coverage with rater agreement. Raters are not informed which items are duplicated. Fleiss' Kappa is calculated on shared items to assess rater consistency. This process follows whether all three raters similarly evaluate a response from a model as being correct or incorrect. Raters are instructed to give a response a score of "1" if it matches or conveys the same key content as that of the dataset's ground-truth answer, and "0" if the information is not decipherable, or is incorrect using the ground truth as the absolute standard. This means that raters have to critically engage with the response when comparing it to the ground truth. The instruction gives raters flexibility when the response does not bear the exact wording from the ground truth found in the dataset.[6]

3.2 Analysis Design

Differences in how the dataset and the LLMs structure their responses make EM a less reliable metric for our evaluation. Ground-truth answers from the SQuAD dataset[7] are typically brief and direct, while LLM outputs tend to be longer and more conversational, reducing the utility of EM.

We also consider the F1-score, but adopt a simpler binary human evaluation scheme, which aligns more directly with our focus on answer correctness (score of 1) or incorrectness (score of 0). This scheme supports direct comparison of performance across dialects, with each SAE question and its variant treated functionally equivalent.

4 Results

A Fleiss' Kappa of 0.889 reflects strong inter-rater agreement. This provides greater confidence in the results for the evaluations that were not shared among raters.

4.1 LLM Performance

Gemini produces the fewest errors, with limited variation across the three dialects. However, despite prompt conditioning, Gemini frequently generates longer-than-expected outputs.

ChatGPT produces the second fewest errors, with SAE performing best, followed by AAVE, and WAPE showing the lowest accuracy.

[2]Clickable link: Mr.Dialect
[3]Clickable link: PidginUNMT
[4]Full examples can be found in Table 3 in Appendix A
[5]Clickable link: Github Repository

[6]See examples in Table 4 in Appendix A
[7]SQuAD is publicly available dataset under CC BY-SA 4.0 license.

Grok performs worst overall,[8] with AAVE slightly better than SAE, while responses to WAPE yield the most errors—up to 70 out of 100. Grok also shows the widest error-rate range.[9]

4.2 Dialect Results

Across the dialects—SAE, AAVE, and WAPE—error rates vary significantly. WAPE shows a marked drop in accuracy (increased error rates) compared to SAE, supporting continued investigation of $\mathbf{RQ_{WAPE}}$ in future work.[10] This contrasts with $\mathbf{RQ_{AAVE}}$, where most LLMs show a minor decrease in accuracy.

To assess these differences, we conduct binomial tests in R (ver. 4.4.3) using binary correctness labels (R Core Team, 2025). For each LLM, we compare the mean SAE error rate to the mean error rate for the corresponding dialect. This enables evaluation of intra-LLM performance: how each LLM handles dialects relative to its SAE baseline.

Table 1 shows binomial test results comparing each LLM's performance on dialect inputs to its SAE baseline. Mean Error Rate reflects the proportion of incorrect answers (out of 100 queries), as rated by three evaluators. The *p-value* indicates whether the dialect error rate differs significantly from the SAE baseline.

LLM	dialect	Mean Error Rate	SAE Mean Error Rate	p
Gemini	AAVE	0.48	0.45	0.331
Gemini	WAPE	0.53	0.45	0.075
ChatGPT	AAVE	0.57	0.54	0.332
ChatGPT	WAPE	0.64	0.54	0.032
Grok	AAVE	0.55	0.56	0.645
Grok	WAPE	0.69	0.56	0.007

Table 1: Binomial Tests comparing dialect errors in LLMs to corresponding SAE performance for same LLM. $N = 100$, $\alpha = .05$.

As an example, Gemini's responses to AAVE queries result in a mean error rate of 0.48. When we compare this value to the SAE mean error rate of 0.45, it yields a non-significant p-value of 0.331. By contrast, ChatGPT's WAPE responses have a mean error rate of 0.64 versus its SAE baseline of 0.54, with a statistically significant p-value of 0.032. With a significance threshold of $\alpha = 0.05$, only ChatGPT and Grok show statistically significant increases in error for WAPE inputs. These findings raise important questions about the sources and implications of LLM errors, which we explore in the following discussion.

5 Discussion

The following observations arise from notable LLM behaviors in response to the provided prompts. These patterns suggest directions for future experimentation and deeper analysis.

5.1 Same Question: Different Answer

A consistent observation is that LLMs often produce different answers to the same query, depending on the dialect used. When inspecting inaccurate responses, LLMs rarely indicate they do not understand the question. Instead, they attempt to respond, with dialect-specific vocabulary or grammar causing misinterpretation. This supports continued exploration of $\mathbf{RQ_{WAPE}}$.

For example, the West African interjection "Chai" is interpreted by an LLM as a scientist's name in a question about NASA. These errors suggest a need for developers to improve model robustness to non-standard varieties of English.

5.2 Responses Mirroring Dialects

Some LLM responses mirror the input dialect in their responses, while others default to SAE. For instance, a prompt written in WAPE may receive a reply in WAPE, but similar prompts in WAPE or AAVE often yield responses in SAE.[11] LLMs are generally designed to be easily understood, which is why responses are typically framed in a conversational format—to aid comprehension. If mirroring the input dialect enhances user comprehension, then LLMs must strive to do so more reliably without explicit prompting.

In short, an important future direction is to evaluate not just correctness, but also whether LLMs adapt stylistically to match user inputs.

6 Conclusion and Future Work

This study explores how LLMs respond to prompts written in AAVE and WAPE compared to SAE. By translating a QA dataset written in SAE into AAVE and WAPE and evaluating LLM performance, we assess how accuracy varies across dialects. Results reveal notable performance gaps—particularly

[8] Error rates for dialects by LLMs are shown in Figure 3, Appendix B

[9] Error rates for all LLMs are in Figure 5, Appendix B

[10] Error rates for dialects are in Figure 4, Appendix B

[11] Fuller examples of questions and responses are provided in Table 5 in Appendix B

WAPE—indicating discrepancies in how LLMs handle non-standard varieties of English.

Given widespread LLM usage, these findings matter, as many users communicate in dialects not well reflected in training data. When LLMs struggle with these variations, they risk misunderstanding entire user communities. Improving performance on dialects like AAVE and WAPE strengthens model generalization across varied inputs.

Several directions for future work emerge from this study's limitations and findings. One priority is to expand the set of dialects and regional varieties tested, which would help determine whether the current findings are generalizable.

Another area for exploration involves distinguishing between a model's understanding of a prompt and the correctness of its response, treating comprehension and accuracy as separate parameters. In some cases, LLMs misunderstand a prompt and respond incorrectly. In others, they appear to grasp the core the meaning of a prompt but still generate an incorrect answer. This distinction is not captured by current metrics, so future work will include an additional evaluation layer to track these two forms of hallucination.

Limitations

This study focuses on only two dialects (AAVE and WAPE), limiting its dialectal scope. It also examines only text-based prompts, excluding other modalities such as speech.

The translation processes differ: AAVE prompts are generated using a public tool, while WAPE translations rely on a trained model. Although reviewed by native speakers, these inconsistencies may affect reliability. Manual scoring by human raters also constrains the study's scale, and technical barriers limited full API access for some LLMs.

Evaluation uses a binary scoring system, which may oversimplify complex outputs—such as clarifying responses or partial answers. Future work will explore more nuanced scoring to better capture multi-turn interactions.

Since the utilized dataset (SQuAD) is several years old at this point, some of the questions randomly have answers that do not reflect current information. This creates a drift between current knowledge and the cataloged ground truth value and can raise the error rate among all questions. Future iterations will look to first filter these questions out from the selection to provide a better sense of the LLM's true error rate.

Although we translate 1,000 questions, we evaluate only 100 in this study. The limitation is not dataset availability but human annotation capacity. While the methodology is designed for replication and scaling, the three raters already score 900 responses. Future iterations will expand the number of questions to enable broader experimentation.

Ethical Considerations

All LLMs are accessed via public APIs under non-commercial research terms, with total usage under $15 USD. Experiments run using newly created accounts and API keys.

Prompt content is drawn from the public SQuAD dataset and contains no personal or sensitive data. No fine-tuning is performed, and all prompts are general, fact-based questions.

While AAVE and WAPE translations are created using different tools with varying transparency, all outputs are reviewed by native speakers. This study is exploratory and not intended to draw prescriptive conclusions.

Finally, some LLM outputs reveal misinterpretation or unintended bias in response to dialectal input. These issues reflect model limitations, not flaws in the dialects themselves. We caution against treating LLM performance as a proxy for language validity.

Acknowledgements

This work would not have been possible without the generous startup support provided by Dr. Herbert Wertheim through the Herbert Wertheim College of Engineering at the University of Florida.

References

David Adelani, Seza Doğruöz, Iyanuoluwa Shode, and Anuoluwapo Aremu. 2025. Does generative ai speak nigerian-pidgin?: Issues about representativeness and bias for multilingualism in llms.

AtlasLS. 2021. English: 3 distinctly different dialects that are spoken in the united states.

Kindra Cooper. 2023. Openai gpt-3: Everything you need to know [updated].

Paresh Dave. 2023. Chatgpt is cutting non-english languages out of the ai revolution.

Nicholas Faraclas. 2017. The survey of pidgin and creole languages.

Abhay Gupta, Ece Yurtseven, Philip Meng, Sean O'Brien, and Kevin Zhu. 2024. Aavenue: Detecting llm biases on nlu tasks in aave via a novel benchmark. In *arxiv*. Algoverse AI Research.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, and Valentin Hofmann. 2016. One language, many gaps: Evaluating dialect fairness and robustness of large language models in reasoning tasks.

Pin-Jie Lin, Muhammed Saeed, Ernie Chang, and Merel Scholman. 2023. Low-resource cross-lingual adaptive training for nigerian pidgin. *arXiv (Cornell University)*.

Kelechi Ogueji and Orevaoghene Ahia. 2019. Pidginunmt: Unsupervised neural machine translation from west african pidgin to english. *Preprint*, arXiv:1912.03444.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Dion Wiggins. 2025. All llms now perform about the same. right? - dion wiggins - medium.

Walt Wolfram. 2020. Urban african american vernacular english.

Kofi Yakpo. 2024. West african pidgin: World language against the grain. *Africa Spectrum*, 59(2):180–203.

A Reproducibility and Evaluation of LLM Models and Question Variations

Tables in this section demonstrate important information such as the specific model versions of the LLMs utilized are noted for reproducibility and the articulated differences in how questions are written, as well as demonstrate a sample of the evaluations that were conducted over the questions.

Company	LLM	Version
Google	Gemini	Gemini 2.0 Flash
OpenAI	ChatGPT	gpt-4o-mini
X	Grok	grok-2-1212

Table 2: Versions of LLMs tested

Dialect	Question
SAE	Along with drought, what is one other factor that is pushing the Amazon rainforest towards a tipping point?
AAVE	Along with drought, what else pushin' the Amazon rainforest toward a tipping point?
WAPE	Along wit drought, wetin na one other factor dat na pushin' di amazon rainforest towards a tippin' point?

Table 3: Example SAE question with WAPE and AAVE equivalents

Question	Ground Truth	LLM Answer	Eval
When did the North American French and Indian War end?	Treaty of Paris 10 Feb 1763	Ended with the signing of the Treaty of Paris in **1763**.	1
When was Luther's last sermon?	15 Feb 1546	Delivered on **Jan 17, 1546**, in Eisleben.	0

Table 4: SAE Questions and LLM responses with correctness evaluation: Correct (1) / Incorrect (0)

B Performance of LLMs for SAE, AAVE, and WAPE

Figures and tables in this section provide detailed information about the performance and error rates of the LLMs used in the experiments.

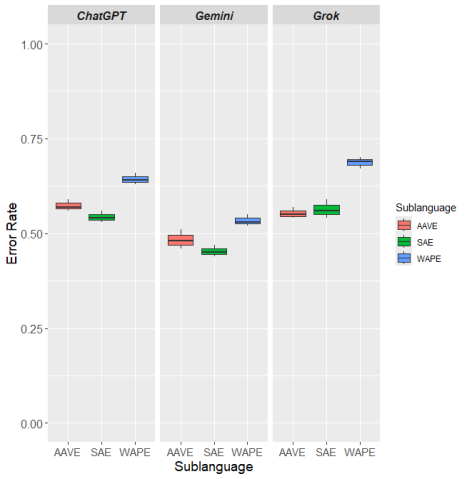

Figure 3: Errors produced across dialects grouped by LLM

Dialect	Question	Abbrev. Response	Type
WAPE	Infrastructure na often called wetin?	Dem dey often call infrastructure “di backbone of society”...	WAPE
WAPE	Hu wrote bout di great pestilence in 1893?	You’re likely thinking of the bubonic plague outbreaks...	SAE
AAVE	Some folks be sayin’ there’s a connection ’tween capitalism, imperialism, and what?	Some folks be sayin’ there’s a connection ’tween capitalism, imperialism, and **colonialism**...	AAVE
WAPE	Chai, hu bin di first nasa scientist in space?	There seems to be some confusion. While Dr. Chiao is the first *American* astronaut of Chinese descent to travel to space...	SAE
WAPE	Abi hu bin kennedy ’ s vice president?	Lyndon B. Johnson war der Vizepräsident von John F. Kennedy. Nachdem Kennedy ...	GER

Table 5: Dialect questions with different response types

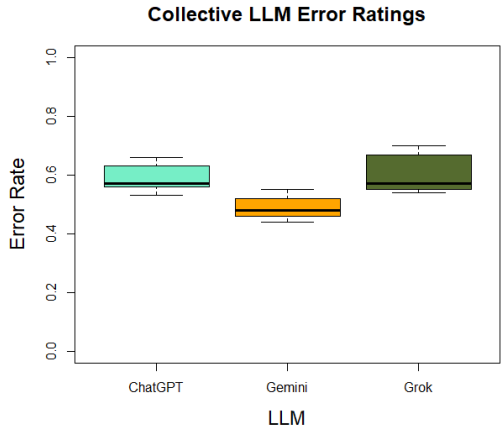

Figure 4: Errors produced across LLMs

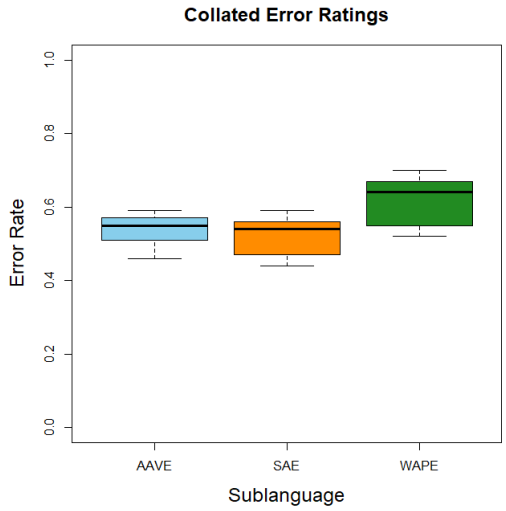

Figure 5: Errors produced across dialects