Evaluating LLMs' Grammatical Comprehension Across Dialects
LING 380-780: Topics in Computational Linguistics: Neural Network Models of Linguistic Structure

Nicole Lam
nicole.lam@yale.edu

Katelyn DeKeersgieter
katelyn.dekeersgieter@yale.edu

1 Introduction

While the use of large language models (LLMs) has become widespread in recent years for daily tasks ranging from next word prediction for iMessage texts to conversing with an Amazon Alexa, blatant disparities in quality of experience using such technology remain. In fact, a recent study found that LLMs demonstrate significant bias and reduced performance when processing African American Vernacular English (AAVE) in reasoning-based tasks compared to processing Standard American English (SAE) (Lin. et al, 2024). This study speaks to the general trend of LLMs' reduced performance on diverse speech-patterns that differ from the SAE used in standard training corpora.

This trend is particularly concerning given that tens of millions of English speakers utilize diverse dialects that are often not represented in training corpora due to prejudicial notions that other dialects are somehow "less grammatical" than Standard American English. While this trend and many other patterns of bias in LLMs' performance have been recorded, there have been few investigations into the exact nature of if and how LLMs fail to comprehend grammatical structures in dialects other than SAE, motivating our project. All code for this project is open-source and readily available at the following link: Github-Dialects.

2 Objectives

Our project aims to broadly investigate whether LLMs have inherent bias against certain dialects of English, specifically African American Vernacular English (AAVE), via reduced grammatical comprehension compared to SAE that could explain the concerning disparity of LLMs' performances across English dialects. To address this question, we propose an analysis of LLM performance on a question answering task for SAE and AAVE, where questions measure grammatical comprehension. This task provides a notion of LLMs' prejudice against AAVE by comparing LLMs' competence for processing grammatical structures in SAE and AAVE. Differences in LLMs' comprehension abilities for SAE and AAVE would reveal that LLMs' understanding of "grammaticality" differs between the two dialects.

The question answering task assesses grammatical comprehension via constituent identification. A linguistic constituent is defined as a word or phrase that functions as a single syntactic unit within a sentence. Examples of constituents include noun phrases, verb phrases, nouns, etc. Thus, parsing sentences via constituent identification constitutes a grammar comprehension mechanism. Linguistics often turn to 4 popular tests to determine constituency: the substitution test, the movement test, the cleft test, and the question-answer test (Anderson, 2018).

This project models our question answering task after the constituency question-answer test, where if a word or string of words is a constituent, "it's usually grammatical for it to stand alone as the answer to a question based on the sentence" (Anderson, 2018). For example, consider the sentence "The students love learning about neural networks". We can determine that "the students" and "neural networks" are constituents via the question-answer test: "Who loves learning about neural networks? The students." and "The students love learning about what? Neural networks." On the other hand, we can determine that "about neural" is not a constituent via the following resulting (ungrammatical) question-answer test: "*The students love learning what networks? About neural.*" Our project adopts this linguistic framework to assess LLMs' grammatical comprehension

of AAVE compared to SAE through a question answering task.

3 Approach & Implementation

To construct our dataset, we generated equivalent pairs of sentences, questions, and constituent answers for Standard American English and African American Vernacular English. We chose 11 subjects: I, you, we, they, he, she, Mary, John, the girls, the boys, and the grandmother. We wrote 100 accompanying subject predicate pairs for the two dialects based on Jack Sidnell's "African American Vernacular English (AAVE) Grammar." Examples of some of our subject predicate pairs for the two dialects include "I lost the game: I never lost no game", "We played with the dog: We be playin' with the dog", "The boys cleaned the room: the boys done cleaned the room.", and "You know her: You be knowin' her".

We created a question and constituent answer to accompany each predicate pair and subject configuration i.e., "Cleaned what? The room: Done cleaned what? The room". We then generated three datasets from these subject-predicate-question-answer combinations: the SAE-Dataset (1100 entries in Standard American English), the AAVE-Dataset (1100 entries in African American Vernacular English), and the Mix-Dataset (all 2200 entries). We utilized CSV formatting for each dataset to delineate the input sentence, question, and target output constituent answer. We also automatized the standard 80-10-10 split to randomize the segmentation of each of the three datasets into training data, validation data, and evaluation data.

After loading in the three datasets (SAE, AAVE, and mixed), we initialized 3 Seq2Seq models with contrasting hyperparameter configurations for recurrent type, bidirectionality, attention, embedding size, hidden size, attention size, batch size, and learning rate. These parameters are summarized in Table (1).

For training these 3 Seq2Seq models, we ran 50 trials of training for each of the 3 training datasets, resulting in 450 total trials.

Parameter	A	B	C
Recurrent Type	gru	rnn	rnn
Bidirectional	True	False	False
Attention	Yes	Yes	Yes
Embedding Size	100	20	75
Hidden Size	100	20	75
Attention Size	50	20	70
Batch Size	5	5	5
Learning Rate	0.01	0.01	0.01

Table 1: Summary of Seq2Seq Model Configurations

4 Evaluation

To first evaluate the highest performing Seq2Seq model configuration, we outputted box plots measuring the 3 Seq2Seq models' validation accuracies across the 50 training trials, demarcating the median validation accuracy, mean validation accuracy, and average test accuracy for each of the three datasets. Then, for each of the three Seq2Seq configurations, we evaluated 3 model iterations (trained on SAE, trained on AAVE, and trained on the mixed dataset) on the SAE test set, the AAVE test set, and the mixed test set (consisting of both SAE and AAVE).

5 Results

As evidenced by our box plots in Figure 1 assessing the three Seq2Seq models' validation and test accuracies across 50 trials for the three datasets, Seq2Seq Model B performed the worst throughout the training process, as shown by the training accuracy curves in Figure 3(b). Seq2Seq Model B had the lowest mean validation accuracy and lowest average test accuracy for all three training modes (Standard American English, AAVE, and the "mixed" pairing of both Standard American English and AAVE). This predicts the evaluation results recorded in Table 3: Seq2Seq Model B, for every training set-up, had the lowest evaluation on the SAE, AAVE and the mixed dataset tests. This performance could likely be a product of Model B's low attention size, embedding size, and hidden size (all equal to 20).

The box plots in Figure 1 reveal that Seq2Seq Model A had the most moderate performance out of the three models. Model A performed quite well during training for both the AAVE dataset and the mixed dataset. On the other hand, Model A struggled on the Standard American English dataset, with a similar mean validation accuracy to Seq2Seq Model B.

However, Model A performed similarly to Model C for all three dataset set-ups, despite Model A having the recurrent type of gru (compared to Model C's recurrent type being rnn). Model A's other parameters, however, are more similar to Model C than Model B i.e., a larger value for embedding size, hidden size, and attention size. Model A's moderate performance during training (as demonstrated by the curves in Figure 3) predicts its evaluation performance as displayed in Table 2.

Across all three training sets, Model A typically had accuracies higher than Model B but lower than Model C for all three training configurations. Model A outperformed Model C four times: on the AAVE test for when the models were trained on Standard American English, and on the SAE, AAVE, and Mix dataset tests for when the models were trained on the mixed SAE and AAVE dataset.

The box plots in Figure 1 and validation accuracy curves in Figure 2 definitively illustrate that Seq2Seq Model C had the most accurate performance throughout training. Moreover, Model C had the most consistent performance across the three training datasets compared to Model A and Model B. Figure 3 further demonstrates the trend of Model C's extremely high accuracy throughout the 50 training trials, providing reasoning for Model C's strong performance on the evaluation test sets shown in Table 4. While Model C was not bidirectional, it still outperformed Model A (bidirectional) with few exceptions, suggesting that bidirectionality is not a critical hyperparameter for this question answering task measuring grammatical comprehension.

Across all three Seq2Seq models, Table 2-4 reveals that the least accurate performances occurred for models trained only on SAE, with a strong disparity between such models' performance on the SAE dataset compared to both the AAVE and mixed datasets. All three models achieved their highest accuracies when trained on the mixed dataset.

When the models were trained on SAE they could not reliably parse inputs in AAVE, as shown by the low accuracies in Table 2-4. When the models were trained on AAVE, however, they could generalize beyond their training and have strong grammatical comprehension for SAE in both the SAE and mixed dataset evaluations. In fact, Models A and B, when trained on AAVE, had higher accuracy on SAE than when these models were trained only on SAE.

Model	Training Set	SAE	AAVE	Mix
A	SAE	0.811	0.612	0.746
	AAVE	0.862	0.929	0.909
	Mix	0.997	0.998	0.995

Table 2: Model A test accuracies on SAE, AAVE, and Mix.

Model	Training Set	SAE	AAVE	Mix
B	SAE	0.224	0.262	0.297
	AAVE	0.296	0.421	0.421
	Mix	0.633	0.628	0.665

Table 3: Model B test accuracies on SAE, AAVE, and Mix.

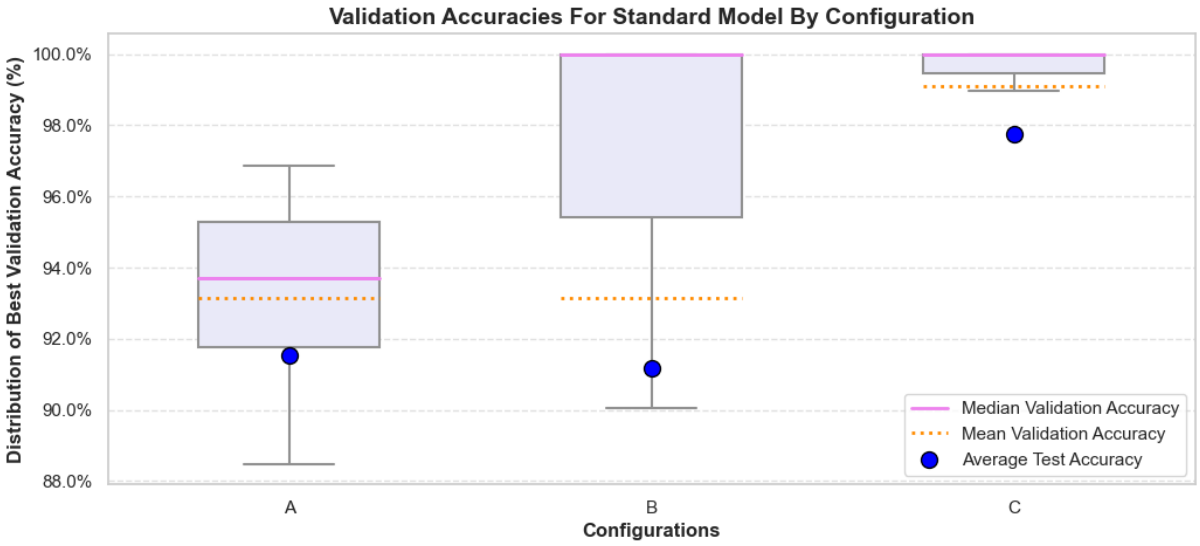
Model	Training Set	SAE	AAVE	Mix
C	SAE	0.842	0.561	0.764
	AAVE	0.781	0.984	0.929
	Mix	0.959	0.995	0.987

Table 4: Model C test accuracies on SAE, AAVE, and Mix.

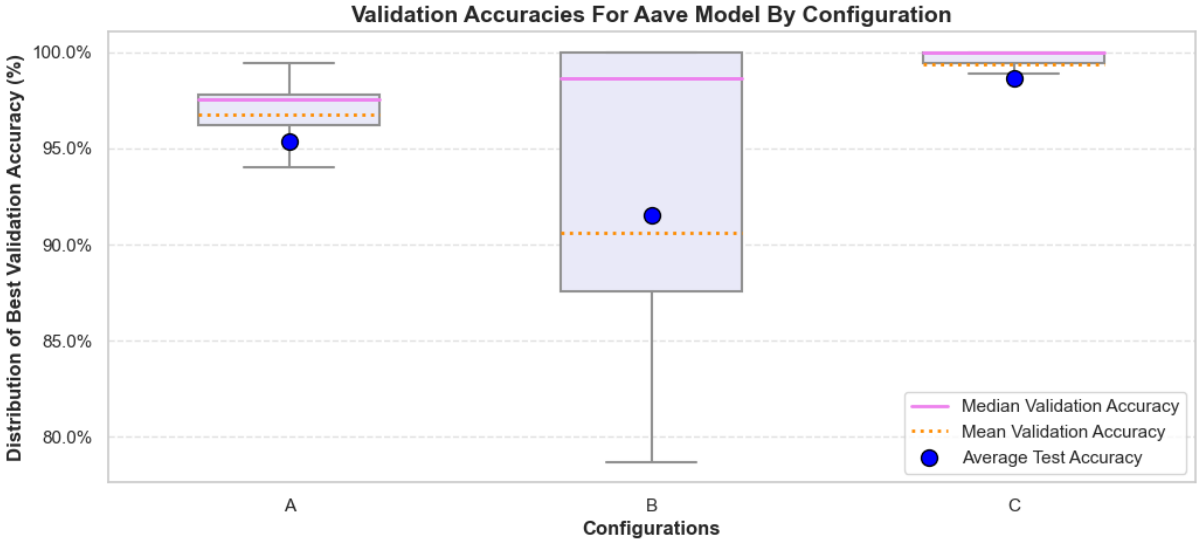
6 Relation to Relevant Literature

This investigation into LLMs' bias against AAVE via reduced grammatical comprehension fits into a larger landscape of studies exploring how LLMs have disparate performances between different dialects of English, impacting LLMs' capabilities as well as users' experience using such technology.

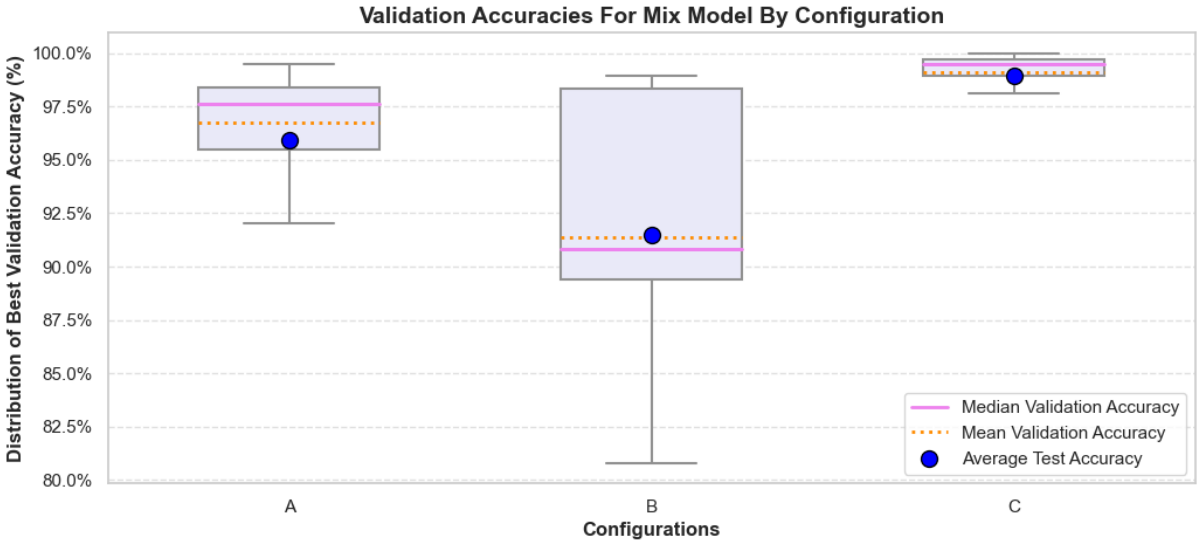
As mentioned in the introduction, Lin et al.'s recent paper aligns with our broad goals. In "One Language, Many Gaps: Evaluating Dialect Fairness and Robustness of Large Language Models in Reasoning Tasks", Lin et al. analyzed how LLMs' performance for canonical reasoning tasks differs when such benchmarks are rewritten using AAVE. The study utilized parallel query pairs in Standardized English and AAVE to evaluate GPT-4o/4/3.5-turbo, LLaMA-3.1/3, Mistral, and Phi-3 (Lin et al., 2024). Lin et al.



(a) Standard Model Validation Accuracy

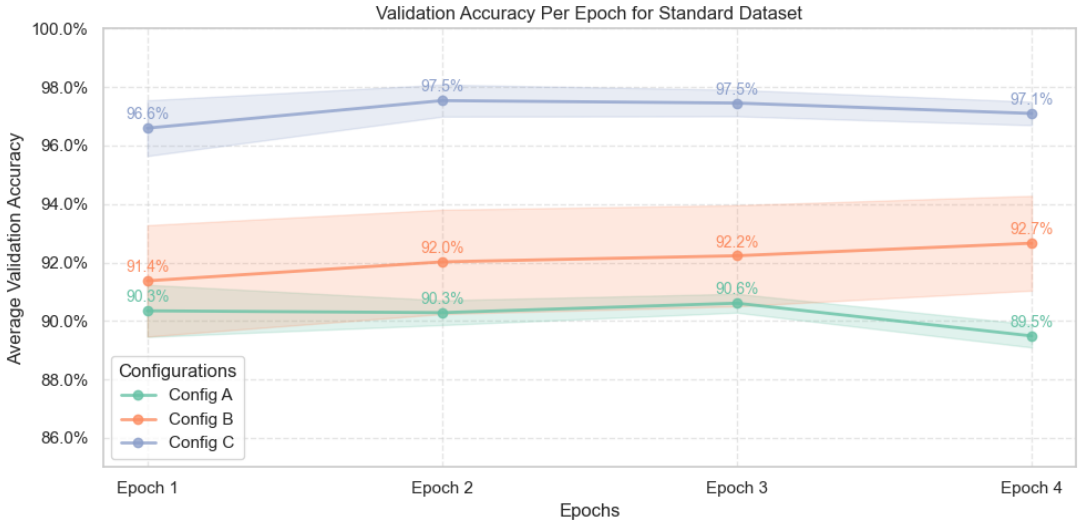


(b) AAVE Model Validation Accuracy

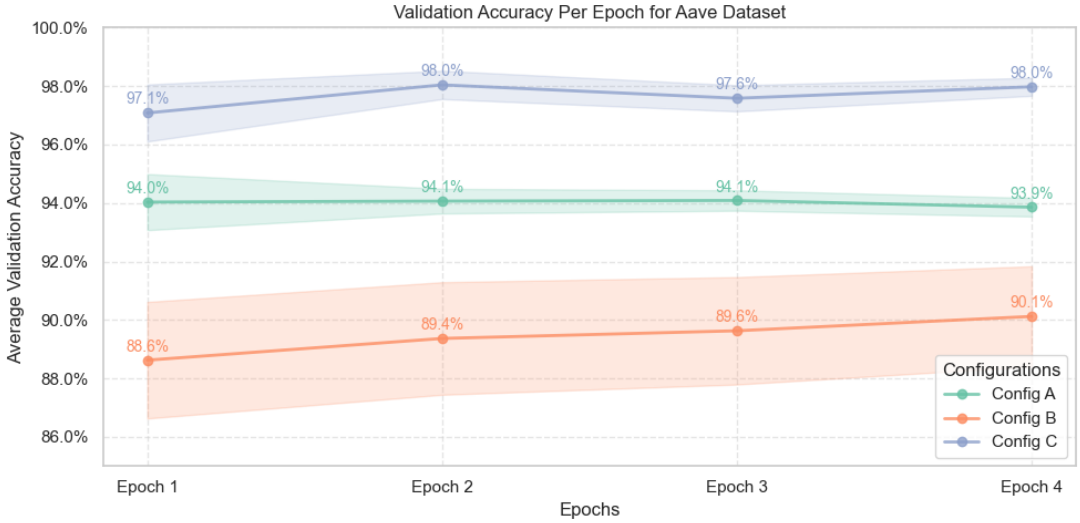


(c) Mix Model Validation Accuracy

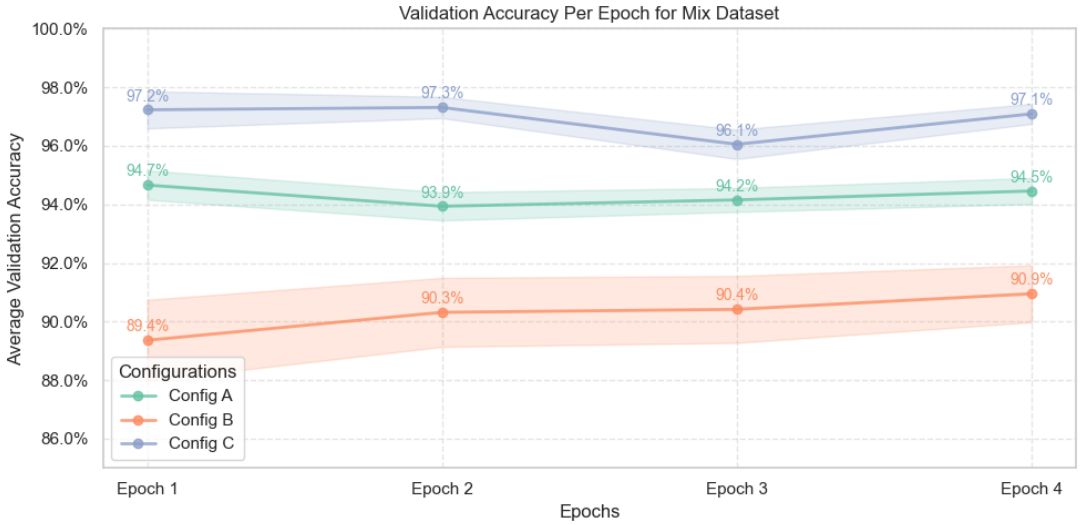
Figure 1: Validation Accuracies for Standard, AAVE, and Mix Models



(a) Standard Model Average Validation Accuracy Per Epoch.

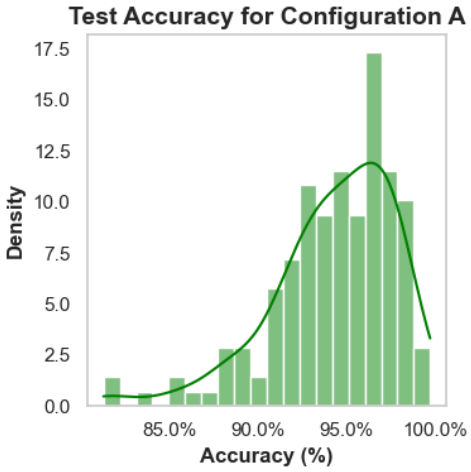


(b) AAVE Model Average Validation Accuracy Per Epoch.

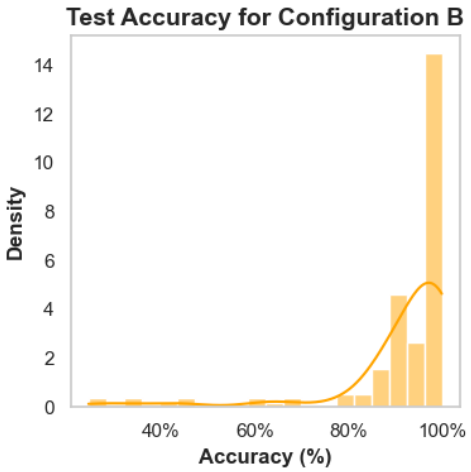


(c) Mix Model Average Validation Accuracy Per Epoch.

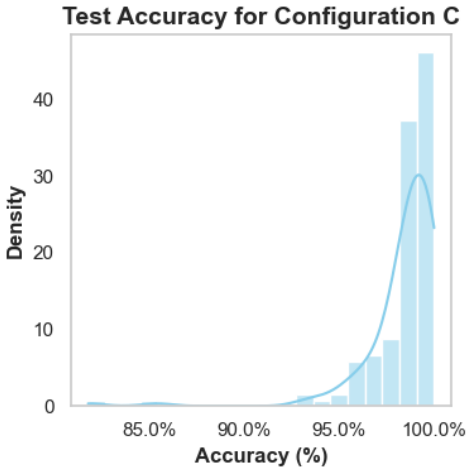
Figure 2: Validation Accuracies per Epoch while training Standard, AAVE, and Mix Models



(a) Configuration A test accuracy distribution over 50 trials of 3 datasets (150 datapoints).



(b) Configuration B test accuracy distribution over 50 trials of 3 datasets (150 datapoints).



(c) Configuration C test accuracy distribution over 50 trials of 3 datasets (150 datapoints).

Figure 3: Test Accuracies while training Standard, AAVE, and Mix Models

found that nearly every model demonstrated "significant brittleness and unfairness to queries in AAVE" and that "AAVE queries can degrade performance more substantially than misspelled texts in Standardized English, even when LLMs are more familiar with the AAVE queries" (Lin et al., 2024). Our project of analyzing inherent disparities in LLMs' syntactical understanding of AAVE underlies the contrasting LLM reasoning performances found by Lin et al.

Similar to Lin et al., Holt et al., in "Perceptions of Language Technology Failures from South Asian English Speakers", aimed to explore how disparities in performance with certain dialects for LLMs impacts use of language technology. Focusing on South Asian Englishes, Holt et al. paired a survey on such issues with an assessment of these issues' pervasiveness in more recently developed LLMs. Such issues assessed included stand-alone dialect words, codeswitching, and register and syntax. Holt et al. completed this assessment by having LLMs select correct definitions for words and codeswitching.

Syntax was evaluated via minimal pairs between Indian English aligned syntax and syntax aligned with Standard American English which were evaluated against equivalent South Asian English constructions that were grammatically incorrect within the dialect, with evaluating which construction (grammatically correct Indian English or grammatically incorrect South Asian English) would be assigned higher probability by the LLM (Holt et al., 2024). Holt et al. concluded that LLM performance dips significantly in the presence of South Asian English lexical & syntactical features in comparison to equivalent Standard American English lexical & syntactical features, a finding that is supported by the parallels our project.

We believe that our project and others like it are necessary to build a more equitable future in the realm of LLMs, as demonstrated by Hofmann et al.'s "AI Generates Covertly Racist Decisions about People Based on their Dialect". Hofmann et al. found that LLMs "embody covert racism in the form of dialect prejudice, exhibiting raciolinguistic stereotypes about speakers of African American English (AAE) that are more negative than any human stereotypes about African Americans ever experimentally recorded" (Hofmann et al., 2024).

This study reiterates the dangers of dialect prejudice: "language models are more likely to suggest that speakers of AAE be assigned less-prestigious jobs, be convicted of crimes and be sentenced to death" (Hofmann et al., 2024). Perhaps Hofmann et al.'s most frightening conclusion is that "current practices of alleviating racial bias in language models, such as human preference alignment, exacerbate the discrepancy between covert and overt stereotypes, by superficially obscuring the racism that language models maintain on a deeper level" (Hofmann et al., 2024). Our project aimed to pinpoint LLMs' racism on such a deeper level of syntactical comprehension of AAVE.

7 Conclusions

This project demonstrates that LLMs without significant exposure to diverse English dialects often fail to effectively comprehend grammatical structures in such dialects. Across varying hyperparameter configurations, all three Seq2Seq models trained only on SAE failed to have acceptable accuracies on the question answering task for sentences in AAVE. The significant disparity in performance between the SAE and AAVE evaluation accuracies for the well-performing Models A and C when trained on SAE demonstrates bias against AAVE via this disparity in grammatical comprehension. Moreover, for all three models, training on solely SAE also failed to produce desirable results for the mixed dataset of SAE and AAVE, with the highest accuracy only being around 75%.

Considering the scarcity of large training corpora for LLMs containing AAVE, we can deduce from this data that LLMs are often biased against AAVE through such disparities in reduced grammatical comprehension (Graves et al., 2024). However, the validation accuracies throughout the 50 trials of training for all three Seq2Seq models across the three datasets (as shown in Figure 2), were not strong predictors of the models' performance on the AAVE dataset evaluation and mixed dataset evaluation. This trend demonstrates the need for metrics of evaluation that incorporate diverse dialects in order to accurately gauge LLMs' performances across dialects. Unfortunately, such evaluation benchmarks are rare, with standard NLP performance benchmarks often only using Standard American English (Ziems et al., 2022).

Therefore, to rectify this pattern of LLMs' bias against AAVE and other dialects, evaluation benchmarks must reflect the diversity of dialects spoken for a language, tying accuracy to LLMs' proficiency across dialects. Moreover, this project demonstrates how the incorporation of multiple dialects not only rectifies any existing bias issues by significantly bridging disparities in performance on tasks for such dialects, but also boosts overall performance across all such dialects, even Standard American English (note the models' high performance on the SAE dataset when trained on the mixed SAE and AAVE dataset in Table 2-4).

Therefore, robust training datasets including diverse dialects of English can boost the performance of LLMs regardless of the input dialect being used. This is further evidenced by the strong performances on the SAE evaluation dataset by Model A and Model C when only trained on AAVE, exemplifying how dialects often share linguistic traits due to their syntactical and semantic inheritances from earlier iterations of a language (Winford, 1998). This trend could also be a result based on our subject-predicate-question-answer equivalent pairs set-up, where every word or phrase in SAE was somewhat represented in the AAVE dataset, but constructions such as double negatives were only present in AAVE, elucidating why there was better inheritance from AAVE to SAE than vice versa. We also note that the mixed dataset was double the size of our SAE and AAVE datasets as a result of being a combination of the two individual dialect datasets. Additionally, the paired nature of the dialect datasets means that some sentences had nearly identical questions/constituent answers across the two dialects, so the mixed dataset's training could include question/answer pairs very similar to pairs in the evaluation set. Both this design and the size difference between the data could be confounding our results, but the conclusion still stands that the incorporation of English dialects beyond SAE into training data will improve LLMs' overall performance, particularly for the tens of millions of English speakers utilizing dialects other than SAE.

8 Future Directions

To expand this project's methods, we could utilize analyzing more Seq2Seq models with increased diversity in hyperparameter configurations. This addition could result in more insights surrounding what hyperparameter set-ups are most optimal for comprehending the grammatical structure of diverse dialects of English. Another continuation of this project is incorporating more dialects of English, such as geographical and regional dialects (i.e., Southern American English) or more global dialects such as Standard British English. Incorporating many different dialects of English would provide this project with a more precise notion of LLMs' bias via disparities in grammatical comprehension.

Additionally, this project would benefit from the next step of rerunning its methods with increased data (i.e., 2,000, 5,000, or even 10,000 entries per dialect). This expansion of the data would not only likely increase accuracy across evaluation on all three datasets, but would also enable the inclusion of more diverse subjects and predicates, expanding our measure of the models' competence for handling more complex syntactical structures in different dialects. Increased data would also standardize the datasets' sizes such that LLMs trained on the mixed dataset wouldn't have an inherent advantage due to the mixed dataset being double the size of the SAE and AAVE datasets. Another related approach would be working with pre-trained models, measuring the models' performance on the evaluation datasets before and after fine-tuning on the AAVE or mixed dataset.

A further expansion of this project could consist of framing our measure of grammatical comprehension as a next word prediction task rather than a question answering task. This means that instead of a Seq2Seq design with the target output of the constituent answer, a model such as GPT-2 would "fill in" the next word in the sequence: "We played with the dog. We played with what? The ___." With such an approach, we could compare the bias (via reduced grammatical comprehension) of GPT-2, BERT (framed as a [MASK] task), and other LLMs. Another future direction for this project would be adopting other tasks for analyzing grammatical comprehension, such as measuring LLMs' abilities to correctly generate syntax trees across dialects.

References

Anderson, C. Essentials of Linguistics. eCampusOntario Pressbooks (2018). https://ecampusontario.pressbooks.pub/essentials-oflinguistics/.

Graves, E., Aswar, S., Desai, R., Nampelli, S., Chakraborty, S., & Hall, T. (2024). AAVE Corpus Generation and Low-Resource Dialect Machine Translation. In COMPASS '24: ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies. ACM, New Delhi, India.

Hofmann, V., Kalluri, P.R., Jurafsky, D. et al. AI Generates Covertly Racist Decisions about People Based on their Dialect. Nature 633, 147–154 (2024). https://doi.org/10.1038/s41586-024-07856-5

Holt, F., Held, W. & Yang, D. Perceptions of Language Technology Failures from South Asian English speakers. In Findings of the Association for Computational Linguistics: ACL 2024 4067–4081 (Association for Computational Linguistics, 2024). https://doi.org/10.18653/v1/2024.findings-acl.241

Lin, F., Mao, S., La Malfa, E. et al. One Language, Many Gaps: Evaluating Dialect Fairness and Robustness of Large Language Models in Reasoning Tasks. arXiv (2024). https://arxiv.org/abs/2410.11005

Sidnell, J. African American Vernacular English (AAVE) Grammar. Outline of AAVE Grammar, Center for Democracy and Technology (2002). https://cdt.org/wp-content/uploads/2017/11/Outline_of_AAVE_grammar__Jack_Sidnell_2002_1_Afr.pdf

Winford, D. On the Origins of African American Vernacular English - A Creolist Perspective: Part II: Linguistic Features. Diachronica, John Benjamins (1998). www.jbe-platform.com/content/journals/10.1075/dia.15.1.05win

Ziems, C., Chen, J., Harris, C., Anderson, J., & Yang, D. (2022). VALUE: Understanding Dialect Disparity in NLU. In Proceedings of the ACL 2022 Main Conference. arXiv:2204.03031v2 [cs.CL]. https://doi.org/10.48550/arXiv.2204.03031