LING 380

Evaluating LLMs' Grammatical Comprehension Across Dialects

By Nicole Lam and Katelyn DeKeersgieter

QUESTION:

Do LLMs have inherent bias against certain dialects of English, specifically African American Vernacular English (AAVE), via reduced grammatical comprehension compared to Standard American English?

QUESTION:

Do LLMs have inherent bias against certain dialects of English, specifically African American Vernacular English (AAVE), via reduced grammatical comprehension compared to Standard American English?

Our approach: a question answering task assessing grammatical comprehension via constituency

Approach

- **1100 SAE-AAVE Equivalent Pairs**
  - **11 subjects**: I, you, we, they, he, she, Mary, John, the girls, the boys, and the grandmother
  - **100 accompanying predicates** for the two dialects, drawing from Jack Sidnell's "African American Vernacular English (AAVE) Grammar."
  - **Question and constituent answer** to accompany each predicate pair and subject configuration
- Initialized 3 **Seq2Seq** models with contrasting hyperparameter configurations

EXAMPLE:

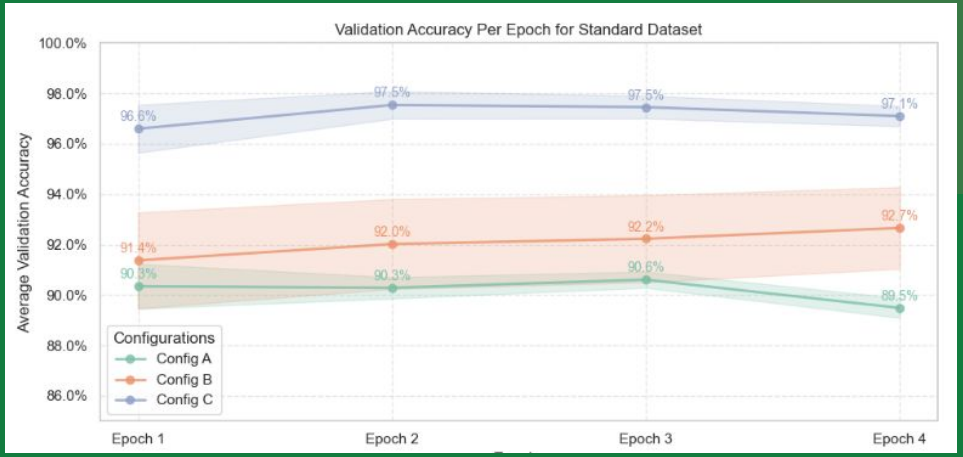
We played with the dog. Played with what? **The dog.**

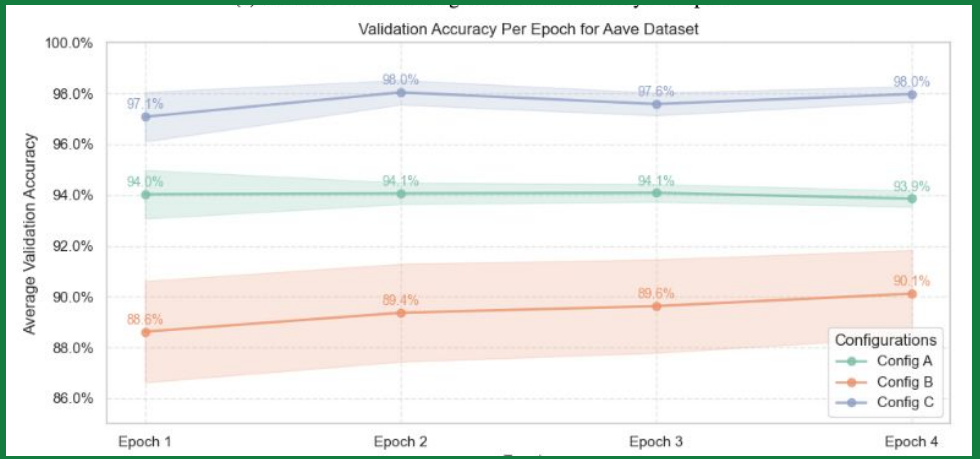
We be playin’ with the dog. Be playin’ with what? **The dog.**

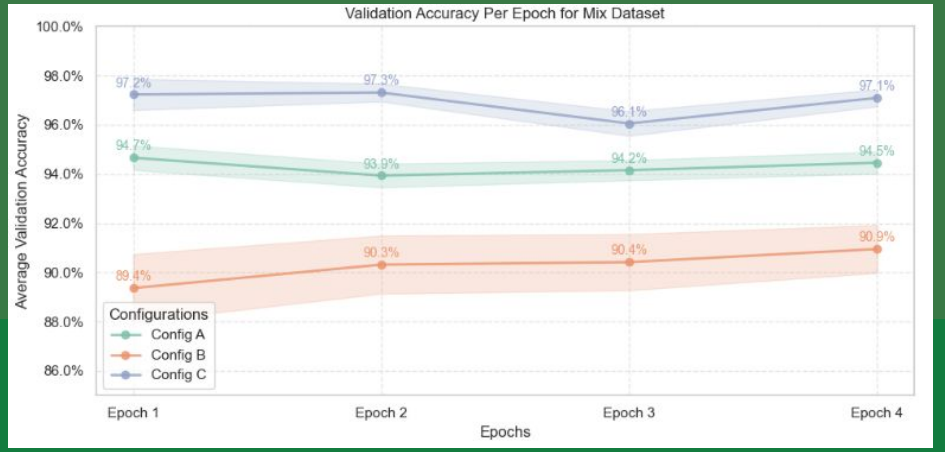
Evaluation

- For all three Seq2Seq model configurations, ran 50 trials of training for each of the 3 training datasets, resulting in **450 total trials!**
- Evaluated each model and training dataset pairing on a Standard American English evaluation dataset, an AAVE evaluation dataset, and a “mixed” dataset consisting of both SAE and AAVE
- Tracked validation accuracy throughout epochs of training and reported final evaluation accuracy across the datasets

RESULTS

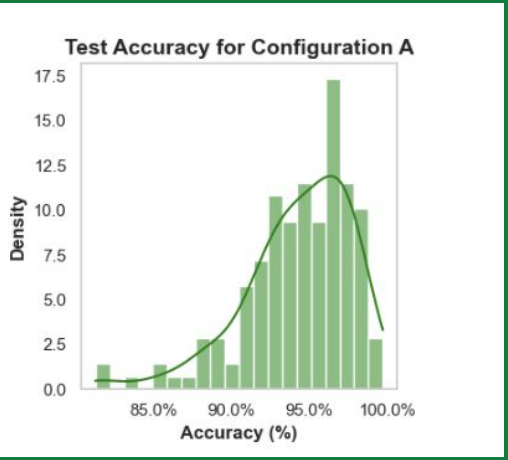


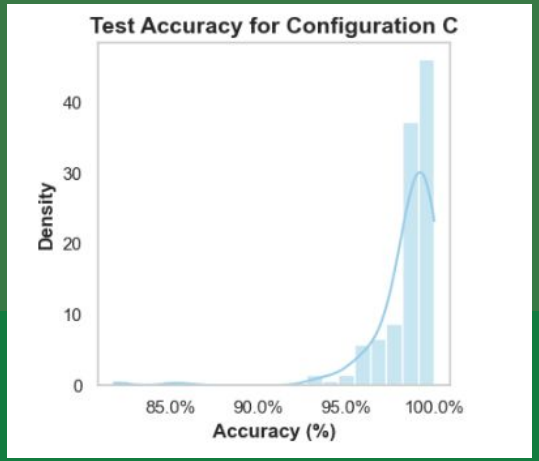


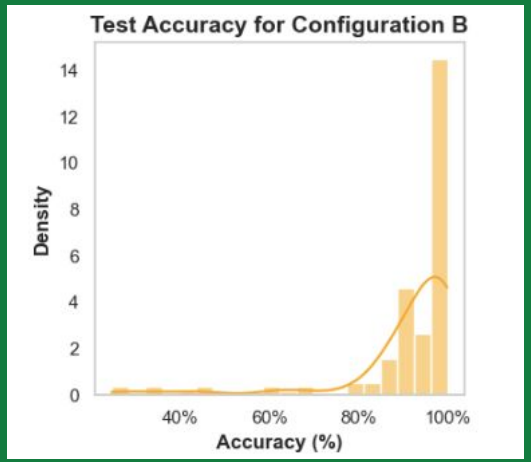


Parameter	A	B	C
Recurrent Type	gru	rnn	rnn
Bidirectional	True	False	False
Attention	Yes	Yes	Yes
Embedding Size	100	20	75
Hidden Size	100	20	75
Attention Size	50	20	70
Batch Size	5	5	5
Learning Rate	0.01	0.01	0.01

RESULTS







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

RESULTS

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

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

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

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

Conclusions

- Models trained only on SAE exhibit **disparity in grammatical comprehension between SAE and AAVE**
- **Multi-dialect training data** boosts overall performance
- Need for evaluation benchmarks that incorporate **dialects beyond SAE**
- Strong performance by models trained only on AAVE → **linguistic inheritance**

Relation to Relevant Literature

Lin et al.: “One Language, Many Gaps”

- Analyzed how LLMs’ performance for canonical reasoning tasks differs when such benchmarks are rewritten using AAVE
- Every model demonstrated "AAVE queries can degrade performance more substantially than misspelled texts in Standardized English, even when LLMs are more familiar with the AAVE queries”

Holt et al.: "Perceptions of Language Technology Failures from South Asian English Speakers"

- Analyzed how LLMs select correct definitions for words & codeswitching for South Asian English Dialects and Standard American English Dialects
- Concluded that LLM performance dips significantly in the presence of South Asian English lexical & syntactical features in comparison to equivalent Standard American features

Hofmann et al.: “AI Generates Covertly Racist Decisions about People Based on their Dialect”

- LMs “embody covert racism in the form of dialect prejudice, exhibiting raciolinguistic stereotypes” about AAVE speakers
- “Current practices of alleviating racial bias in language models... exacerbate the discrepancy between covert and overt stereotypes, by superficially obscuring the racism that language models maintain on a deeper level”

Future Directions

- Increased **hyperparameter configurations**, **dialects**, and **data**
- Framing our measure of grammatical comprehension as a **next word prediction task** rather than a question answering task
  - “We played with the dog. We played with what? The ___.”
- Adopting other tasks for analyzing grammatical comprehension, such as measuring LLMs’ abilities to **correctly generate syntax trees across dialects**

QUESTIONS?