Analyzing Dialectal Biases in LLMs for Knowledge and Reasoning Benchmarks

Eileen Pan[1], Anna Seo Gyeong Choi[1], Maartje ter Hoeve[2], Skyler Seto[2], Allison Koenecke[1,3]
1: Cornell University, 2: Apple, 3: Cornell Tech

LLMs perform up to 20% worse on non-"standard" English dialects.
Just three grammar rules can explain 64-85% of degradation.

Motivation & Problem

- Users write in non-"standard" dialects (e.g., using African American English grammar)
  - LLMs can be unreliable for such users
- We audit the performance of LLMs in answering multiple choice benchmark data in various dialects.
- We investigate: which specific grammar rules drive underperformance?

Methods

🔧 **Multi-VALUE package: Translates Standard American English → dialects**

- Apply grammatical rule transformations to QA dataset questions
- Rules from eWAVE linguistic database for each dialect

📊 **3 QA Datasets**
- BoolQ (9.4K)
- SciQ (11.7K)
- MMLU (14K)

🗣️ **6 English Dialects Audited**
- African American, Appalachian
- Chicano, Indian
- Singaporean, Southern

🤖 **3 LLMs**
- Gemma-2B
- Mistral-7B
- GPT-4o-mini

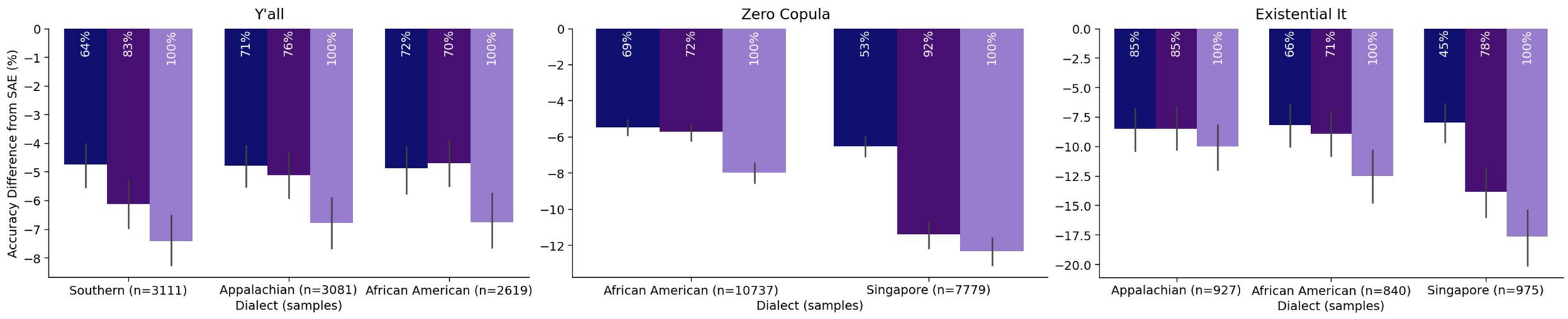
Results

RQ1: Do LLMs underperform on multiple choice questions for dialectal variants? Yes: up to a 20% accuracy drop.

English Variety	BoolQ Accuracy (%)			SciQ Accuracy (%)			MMLU Accuracy (%)		
	Gemma 2B	Mistral 7B	GPT4o-mini	Gemma 2B	Mistral 7B	GPT4o-mini	Gemma 2B	Mistral 7B	GPT4o-mini
Standard American English	100	100	100	100	100	100	100	100	100
Chicano English	93.9 (-6.1)	95.6 (-4.4)	96.7 (-3.3)	99.2 (-0.8)	99.6 (-0.4)	99.5 (-0.5)	89.3 (-10.7)	92.9 (-7.1)	95.2 (-4.8)
Appalachian English	92.0 (-8.0)	93.6 (-6.4)	94.8 (-5.2)	98.1 (-1.9)	99.0 (-1.0)	99.2 (-0.8)	86.8 (-13.2)	93.0 (-7.0)	93.8 (-6.2)
Southern English	90.1 (-9.9)	93.1 (-6.9)	94.8 (-5.2)	98.4 (-1.6)	99.1 (-0.9)	98.9 (-1.1)	83.1 (-16.9)	92.6 (-7.4)	92.4 (-7.6)
African American English	85.9 (-14.1)	91.9 (-8.1)	95.0 (-5.0)	98.2 (-1.8)	99.1 (-0.9)	98.8 (-1.2)	84.4 (-15.6)	92.3 (-7.7)	92.3 (-7.7)
Indian English	86.9 (-13.1)	90.2 (-9.8)	93.6 (-6.4)	97.5 (-2.5)	98.4 (-1.6)	98.5 (-1.5)	81.3 (-18.7)	91.2 (-8.8)	90.8 (-9.2)
Singaporean English	83.3 (-16.7)	88.2 (-11.8)	92.3 (-7.7)	96.4 (-3.6)	98.0 (-2.0)	97.4 (-2.6)	78.4 (-21.6)	89.9 (-10.1)	88.8 (-11.2)

RQ2: Can we decompose this degradation by grammatical rules? Yes: 3 rules explain majority of degradation.

Grammar Rule	English Dialects Occuring In	Example (Standard American English)	Example (with Grammar Rule Applied)
Existential "it"	Appalachian, African American, Singaporean	How many kcal are there in one gram of ethanol?	How many kcal is it in one gram of ethanol?
Zero Copula	African American, Singaporean	Alpha emission is a type of what?	Alpha emission a type of what?
Y'all	Southern, Appalachian, African American	Can you drive with a beer in Texas?	Can y'all drive with a beer in Texas?



Key Findings

- Dialectal biases persist even in basic multiple choice tasks.
- LLMs perform worst on Singaporean English.
- Three grammatical structures drive the most degradation: *Existential "it"* (instead of "there"), *Zero copula* (dropping "be"), and *Y'all* (second person plural).
- Of the grammar rules with highest impact on LLM performance, 9 of the top 20 rules appear across multiple dialects – improving LLM performance on these rules could have an outsized positive impact

Questions?

Link to paper

Link to code

koenecke@cornell.edu