Analyzing Dialectical Biases in LLMs for Knowledge and Reasoning Benchmarks

Eileen Pan, Anna Seo Gyeong Choi,
Maartje ter Hoeve, Skyler Seto, Allison Koenecke
EMNLP 2025 Findings

Motivation & Problem

The Problem:

- Users writing in non-“standard” dialects (e.g., African American English) are underrepresented in training data
- These users suffer from worse LLM responses
- Critical implications for high-stakes scenarios: hiring, criminal justice, education

Research Gap:

- Prior work studied either individual grammatical rules OR overall dialects (Ziems et al., 2023; Srirag et al., 2025)
- Our question: Which specific grammar rules drive underperformance?

Research Questions

RQ1: Do LLMs underperform on multiple choice questions typed in written dialects versus Standard American English?

RQ2: Can we decompose this degradation by specific grammatical rules?

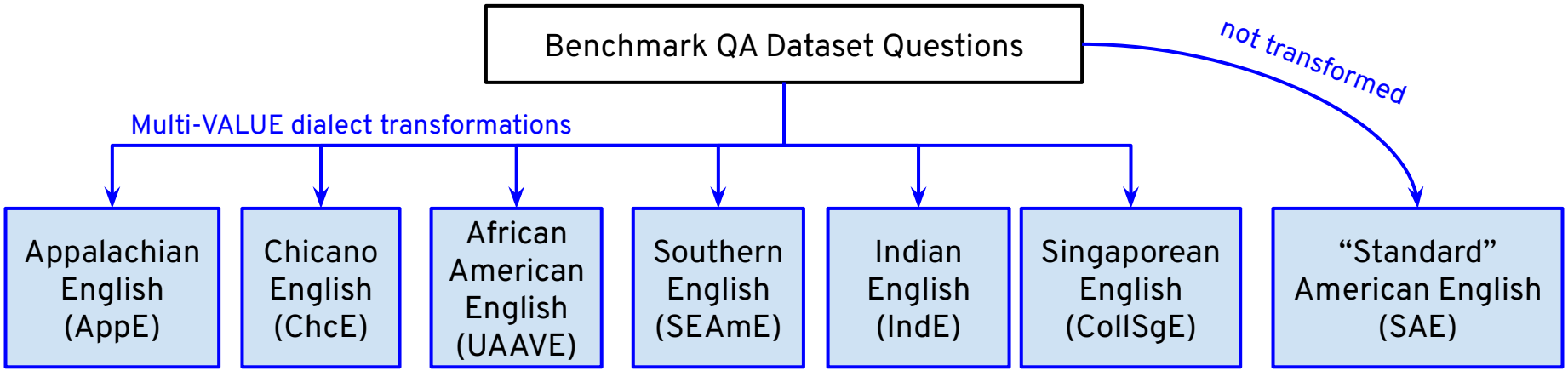
Why this matters: Identifying high-impact grammatical rules can inform targeted model improvements across multiple dialects through transfer learning

Methods

- Auditing with:
  - **3 QA Benchmarks:** BoolQ (9.4K), SciQ (11.7K), MMLU (14K)
  - **3 LLMs:** Gemma-2B, Mistral-7B, GPT-4o-mini
  - **6 English Dialects:** African American, Appalachian, Chicano, Indian, Singaporean, Southern

- Multi-VALUE Package:
  - Transforms Standard American English (SAE) → dialect variants
  - Can apply full dialects OR individual grammar rules

Methods – Full Dialects



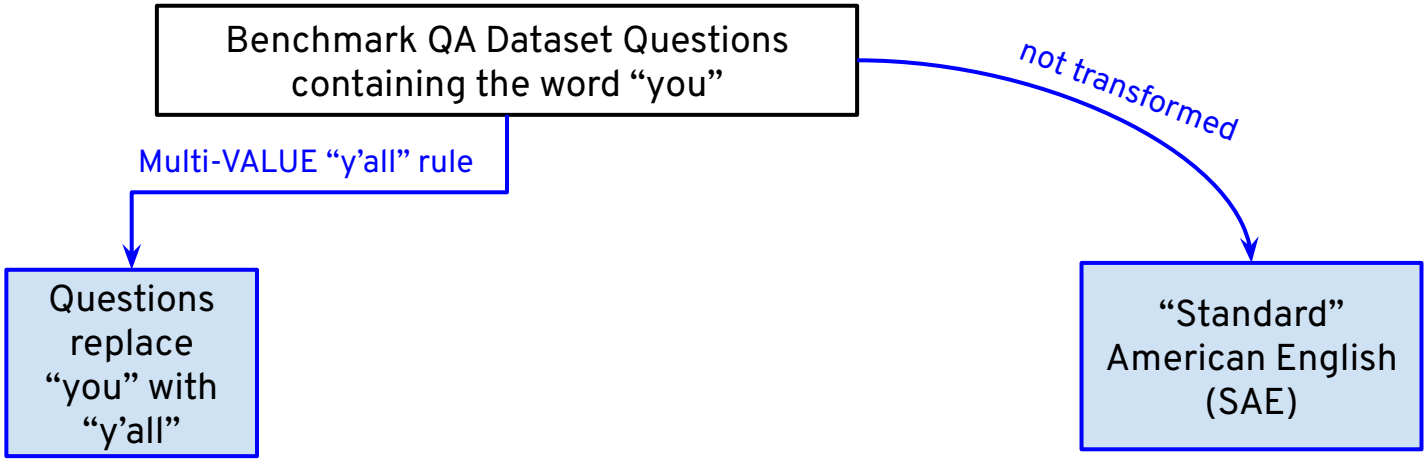
Slide adapted from Emma Harvey, https://dl.acm.org/doi/10.1145/3715275.3732137

RQ1 Results – Dialectal Performance Degradation

- **All** dialects show performance degradation across all tasks, up to ~20 pp
  - Gemma performs 21.66 pp worse on MMLU in Singaporean English

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)

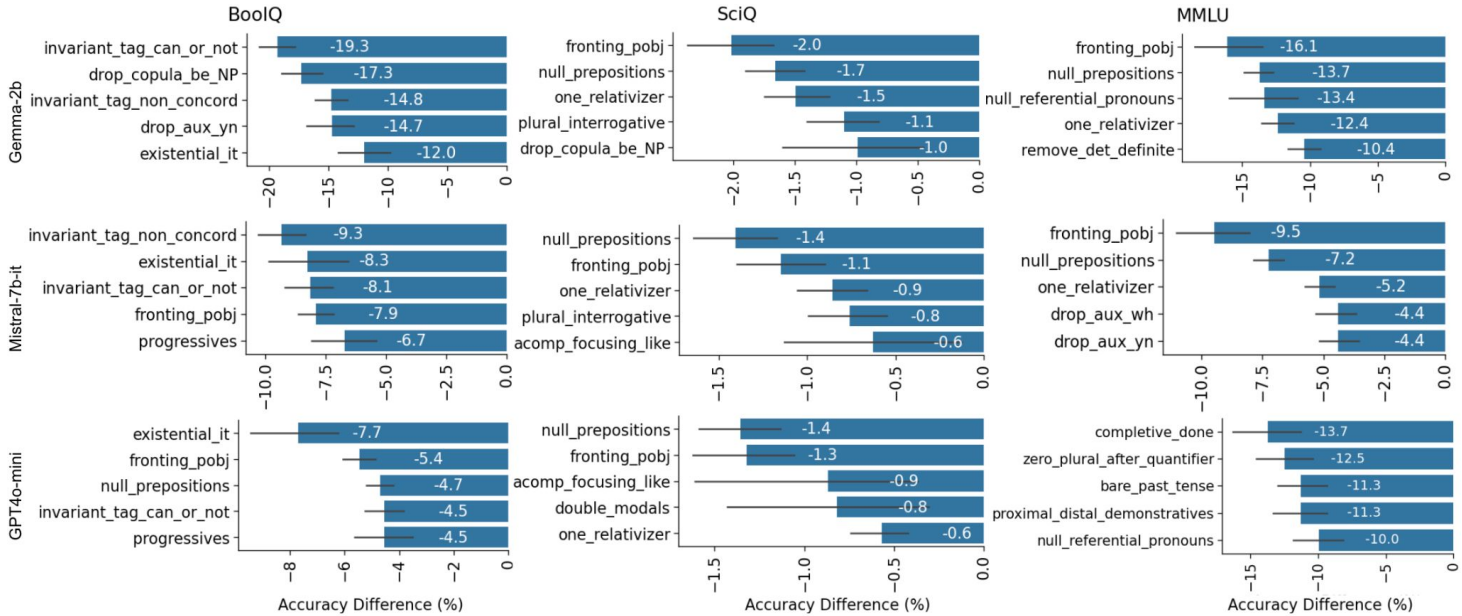
Methods – Individual Grammar Rules



Slide adapted from Emma Harvey, https://dl.acm.org/doi/10.1145/3715275.3732137

RQ2 Results – Individual Grammar Rules

- Different rules cause different impacts across tasks and models

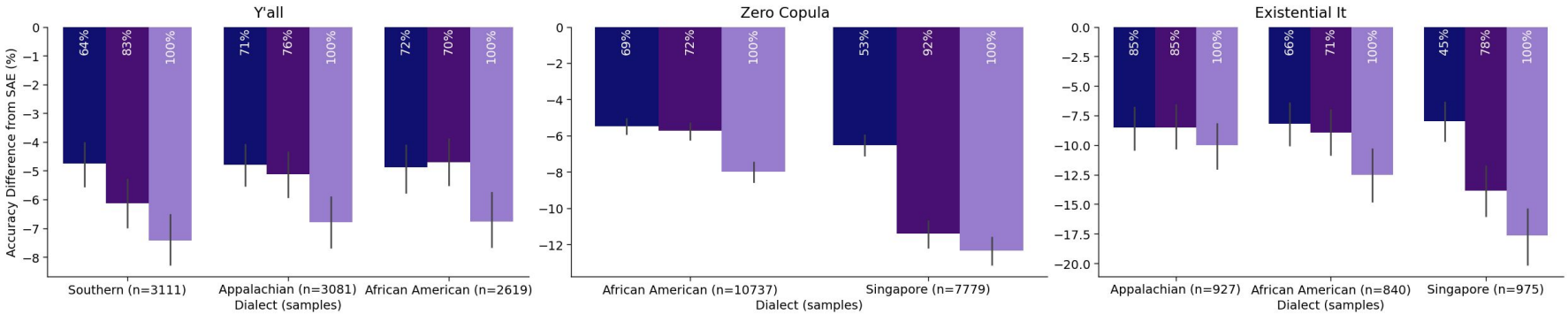


RQ2 Results – Individual Grammar Rules

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?

RQ2 Results – High-Impact Rules Within Dialects

- For dialects where these rules occur:
  One of these three rules account for 64-85% of total dialect degradation



Conclusion

- LLMs show significant dialectal biases even on simple multiple choice tasks
- Three grammar rules (existential it, zero copula, y’all) are high-impact for American English dialects
    - Single rules explain 64-85% of degradation within their respective dialects
- Focused training on high-impact rules could improve fairness across multiple dialects (Held et al., 2023)

Thank You!

References

Held, W., Ziems, C., & Yang, D. (2023, July). TADA: Task Agnostic Dialect Adapters for English. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 813-824). https://arxiv.org/pdf/2305.16651

Srirag, D., Sahoo, N. R., & Joshi, A. (2025, January). Evaluating Dialect Robustness of Language Models via Conversation Understanding. In Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation (pp. 24-38). https://arxiv.org/pdf/2405.05688?

Ziems, C., Held, W., Yang, J., Dhamala, J., Gupta, R., & Yang, D. (2023, July). Multi-VALUE: A Framework for Cross-Dialectal English NLP. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 744-768). https://arxiv.org/pdf/2212.08011