Mitigating Dialect Biases in Privacy Policy Question-Answering Systems through Collaborative Agent-based Language Models

Astrid Bernaga, Đorđe Klisura, Anna Karen Gárate, Rajesh Roshan and Anthony Rios
Master of Applied Artificial Intelligence – Tecnológico de Monterrey,
School of Data Science – The University of Texas at San Antonio,

Project Goal

This is a project to question the **performance disparities** in privacy policy question-answering systems **across different English Dialects** and to create a **collaborative multiagent-based** solution to mitigate these biases, promoting fairness in automated systems.

Introduction

What is the impact of dialect biases in LLMs, particularly in privacy policy QA, and why is it important?

Biased systems: **Increase disparities in language understanding**, affecting marginalized communities, leading to **unequal access** to **critical information**.

Our motivation rises from the need to address these biases, ensuring that all users, regardless of their linguistic background, receive accurate information. **Employing Human-Centered Design (HCD)** principles to create a solution that identifies and mitigates these biases, where experts (a privacy policy LLM) work with community representatives (dialect-specific LLM in our case)

Using a subset from the **PrivacyPolicyQA dataset** with 10922 examples, from which 5549 Relevant and 5474 Irrelevant labels were evaluated with our collaboration multi-agent solution that combines a dialect agent with a privacy policy expert agent to address gaps in fairness and accuracy.

Methods

Step 1. Dataset Generation

We compiled a dataset of privacy policy questions (PrivacyPolicyQA) translated into 50 different dialects, including Aboriginal English, Chicano English, African American Vernacular English, etc.; using Multi-VALUE framework.

Dialect	Question
Standard American English	will you sell my information?
Aboriginal	gon y'all sell me information??
African American Vernacular	might will y'all sell my informations?

Table 1. Example questions in different dialects

Step 2. Multi-Agent Collaboration

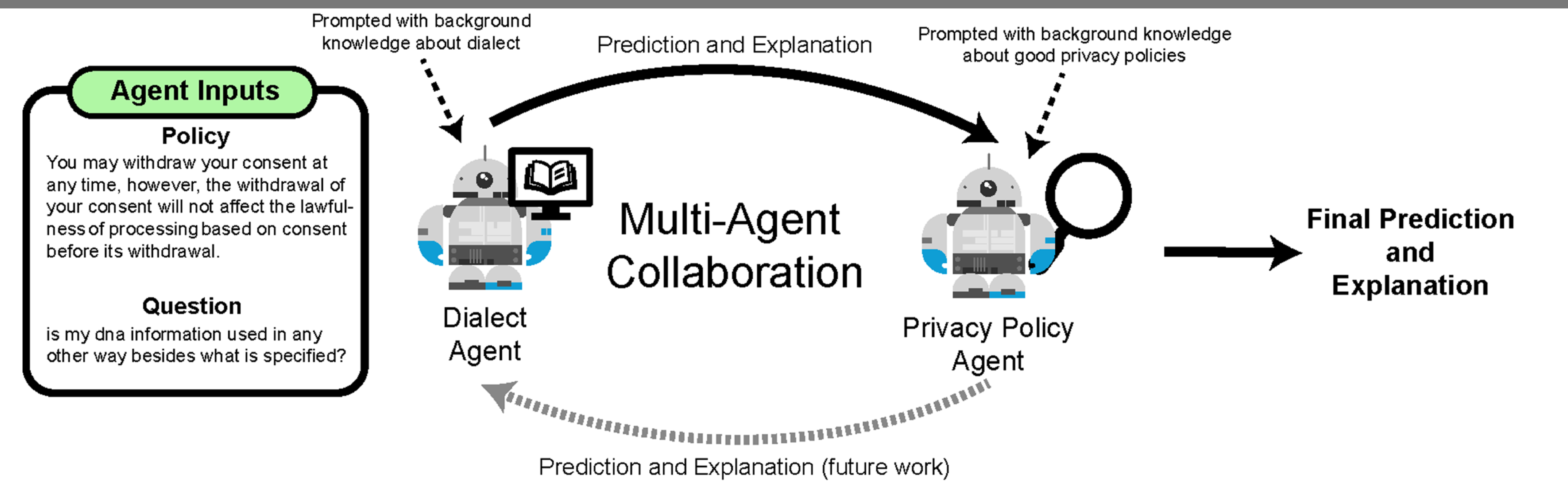
We implemented a multi-agent solution using GPT 3.5-turbo, with two agents: a **Dialect Agent** and **Privacy Policy Agent**.

Step 2a Dialect Agent. The **Dialect Agent** is prompted with a **background about the dialect** (e.g., grammar patterns) a **privacy policy segment** and an **question** in a specific dialect, tasked with translating the text to Standard American English, explaining the relevance of the policy segment to the question, and labeling the answer as "Relevant" or "Irrelevant"

Step 2b Privacy Policy Agent The expert agent is prompted with **background about what privacy policies contain** and what makes a good policy, the **dialect agent's explanation** and **prediction**, the original **question**, **privacy policy segment**. The expert must review the dialects agent's reasoning, verify the label and provide a final label and explanation.

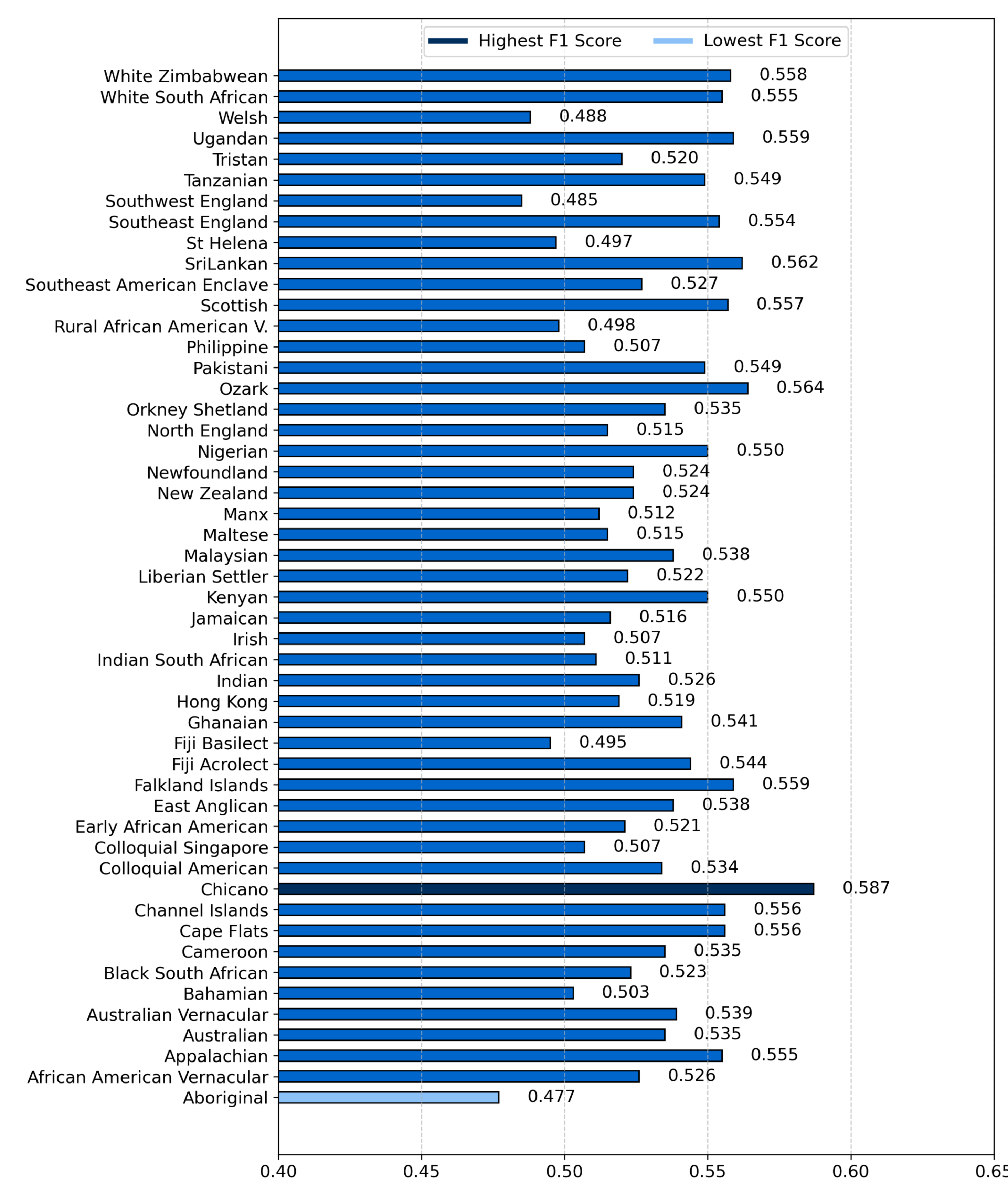
The University of Texas at San Antonio
UTSA School of Data Science

Methodology Overview



Results

RQ1: Does model performance significantly vary across different dialects of English? **YES**



F1 scores from the Baseline model (Few-shot prompting)

RQ2: Can multi-agent collaboration mitigate performance disparities? **YES**

	Aboriginal English	Chicano English	Standard American English	Max Difference	Average Difference
Few-shot Prompting	0.48	0.59	0.59	0.11	0.058
Dialect Agent Only	0.61	**0.67**	**0.66**	0.05	0.03
Multi-Agent Collaboration	**0.64**	**0.67**	**0.66**	**0.02**	**0.015**

We compare to **two baselines**: Few-shot Prompting and Dialect Agent Only

The F1 score for **Aboriginal English** improved from 0.48 to 0.64, and for **Chicano English**, from 0.59 to 0.67. **Standard American English** improved from 0.59 to 0.66. The maximum difference **decreased** from 0.11 to 0.02, and the average difference reduced from 0.058 to 0.015.

These improvements indicate the collaborative model agent incorporating the dialect and expert in privacy policies enhanced the performance across dialects.

Conclusion

There are performance disparities across dialects. Multi-agent collaboration mitigates these disparities.

Future Work

Enhance the multi-agent collaboration system by introducing a **feedback loop** where the **Privacy Policy** Agent and the **Dialect** Agent continually learn from each other's predictions and explanations, incorporating more Dialects and evaluate other LLMs.

Apply the developed methodology to other critical areas like **healthcare** to handle **dialect biases** and improve performance.

References

- Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. 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), pages 744–768, Toronto, Canada. Association for Computational Linguistics.
- Ravichander, Abhilasha, Alan W. Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. "Question Answering for Privacy Policies: Combining Computational and Legal Perspectives." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4947-4958. 2019.

Tecnológico de Monterrey