Diversity Track: Bridging the Linguistic Divide: Evaluating and Enhancing Large Language Models (LLMs) for Code-mixed Language Processing

- Amulya Yadav, College of Information Sciences and Technology
- Rebecca Passonneau, College of Engineering
- Ritu Jayakar, College of the Liberal Arts

This project is jointly funded by the Institute of Computational and Data Sciences.

Abstract: Language is a profound tool for communication and interaction, yet the vast linguistic diversity across the globe poses significant challenges for emerging Natural Language Processing (NLP) technologies. Over the past one year, Large Language Models (LLMs) such as OpenAI ChatGPT and Google Bard have shown remarkable promise in understanding and generating human-like text. However, preliminary research done by the PI suggests that their potential seems skewed towards monolingual English-speaking users, leaving a large portion of the global population—those engaging in code-mixed language usage—at a comparative disadvantage. Code-mixing, the fluid alternation between languages within a conversation or text, is a common linguistic phenomenon, especially in multilingual societies in South Asia (e.g., India) and East Africa (e.g., Kenya). This proposal aims to answer the following research question: How well can state-of-the-art LLMs handle code-mixed data, and how does their performance compare to Natural Language Processing (NLP) methods that are explicitly designed to handle code-mixed text?