Code-Switching in Medical Conversations: Challenges and Opportunities for Indic Languages

B.Tech Project Research Proposal
Nikhil Deepak & Saiuditi Rout

Summary of the Proposal

India's multilingual society often leads to code-switching, where speakers mix two or more languages in the same conversation. This is especially common in medical interactions, where patients and doctors may alternate between regional languages and English medical terminology. Existing language models are mostly designed for monolingual, standardized text and perform poorly when faced with code-switched input, creating barriers to effective healthcare communication. This project aims to study code-switching in Indic medical conversations by collecting and analyzing dialogue text, identifying common linguistic patterns, and benchmarking current language models. The research will further explore methods such as fine-tuning and prompting strategies to improve performance, contributing both datasets and benchmarks for future work.

Background

India is one of the most multilingual societies in the world. People often speak multiple languages and mix them naturally within the same conversation. For example, in many urban areas, Hindi and English are frequently combined ("Hinglish") even in formal contexts. This practice, known as code-switching, is also widespread in regional settings where local languages and dialects blend with English or other regional tongues.

In healthcare interactions, code-switching plays an even more critical role. Patients may describe their condition in their mother tongue but use English words for medical terms, while doctors may switch languages when explaining diagnoses or treatment. Yet most NLP systems, and especially current language models, are built for monolingual input. They often fail to interpret code-switched text correctly, leading to errors in downstream applications such as summarization or dialogue understanding.

The absence of systematic research and resources on Indic medical code-switching creates a significant gap. Studying this problem in the context of large language models can directly improve healthcare technologies for millions of speakers across India.

Goal and Objectives

The project has three main goals:

1. **Linguistic Analysis:** Identify and describe the common patterns of code-switching in Indic medical conversations.
2. **System Evaluation:** Benchmark how well current language models handle code-switched medical dialogue.
3. **Resource/Method Development:** Develop datasets, benchmarks, or model adaptation strategies to improve language model performance.

This leads to the following research questions:

- What kinds of code-switching patterns occur in medical conversations in Indic languages?
- How do current language models perform on these inputs?

- What approaches (data curation, fine-tuning, prompting strategies) can improve language model performance in this domain?

Literature Review

Previous research on code-switching has largely focused on well-studied pairs such as Spanish–English and Mandarin–English, with extensive work on syntactic and sociolinguistic aspects. In the Indian context, some studies have looked at Hindi–English code-switching, but work on regional Indic languages remains very limited.

On the computational side, research on large language models has advanced rapidly, with applications in translation, dialogue summarization systems, and domain-specific text processing. However, nearly all systems assume monolingual, standardized text, most often in English. Very few studies have examined how LMs process code-switched input, and almost none have considered the healthcare domain in Indic languages.

This shows a clear research gap: the intersection of code-switching, Indic languages, and language model performance in healthcare contexts is largely unexplored.

Proposed Work

The proposed research will proceed in three stages:

1. **Data Collection**
    - Use medical conversion datasets from Pranik if available.
    - Collect or simulate doctor–patient dialogue text in multiple Indic languages where code-switching occurs naturally.
    - Annotate the text with language boundaries and medical terminology.
2. **Linguistic and Computational Analysis**
    - Analyze code-switching frequency, triggers, and structures in medical dialogue.
    - Benchmark existing large language models on the dataset.
3. **Exploratory Methods for Improvement**
    - Experiment with fine-tuning or domain adaptation of LMs using the collected dataset.
    - Explore prompting strategies for better handling of code-switched medical dialogue.
    - Propose a benchmark dataset and guidelines for future LM research in this space.