Project Overview

As part of this internship assignment, you will analyze a few conversations between debt collection agents and the borrowers. The goal is to evaluate compliance, professionalism, and call metrics. Each conversation is stored in a YAML file format with detailed utterance-level information.

Data Structure

The name of each file represents the call id.

Each YAML file (in All_Conversations.zip) contains the following fields:

- **speaker**: Identifies whether the utterance is from an agent or borrower
- **text**: Contains the actual speech content
- **stime**: Records the start timestamp of the utterance
- **etime**: Records the end timestamp of the utterance

Analysis Tasks

Question-1. Profanity Detection
  a. Identify all the call ids where collection agents have used profane language.
  b. Identify all the call ids where borrowers have used profane language.

Question-2. Privacy and Compliance Analysis
  a. Identify all the call ids where agents have shared sensitive information (balance or account details) without the identity verification(i.e. without verification of date of birth or address or Social Security Number).

Question-3. Call Quality Metrics Analysis
  a. Calculate overtalk (i.e. simultaneous speaking) percentage per call.
  b. Calculate silence percentage per call.

Implementation Requirements

- **For Question-1 and Question-2:**
  - **Pattern Matching Approach**: Implement regex-based detection systems (Reference).
  - **Machine Learning Approach:** Choose one of the following:
    - Develop a classification model. Note, this approach will require you to annotate the data yourself.
    - Implement a fine-tuned LLM prompt system.
  - **Comparative Analysis Requirements**: Compare the pattern matching and selected machine learning approach to recommend which is the better approach for each of the scenarios.

- **For Question-3:**
    - **Visualization Requirements:** Create visual representations of silence and overtalk metrics.

Deliverables

1. **GitHub Repository Must include:**
    a. Well-documented codes for both the approaches (valid for Question-1 and Question-2).
    b. Visualization code (valid for Question-3).
    c. README file with setup and execution instructions (if required).
2. **Technical Report Must include:**
    a. Implementation recommendations for the different scenarios (valid for Question-1 and Question-2)
    b. Visualization analysis (valid for Question-3)