Milestone 2: Dataset Preparation

1. Dataset 1

Dataset 1

- **Location and Name:**
  Medical Speech, Transcription, and Intent Dataset
  Path: /kaggle/input/medical-speech-transcription-and-intent/Medical Speech, Transcription, and Intent/recordings
- **Description:**
  This dataset contains medical speech recordings and their corresponding transcriptions.
  It is divided into three folders — train, test, and validate.
  Each entry in the provided CSV file (overview-of-recordings.csv) lists an audio file name and its spoken phrase (transcript).
- **Data Preparation:**
  The overview CSV was loaded, and each audio file listed in it was searched within the train, test, and validate folders.
  The complete file paths were then added to a new DataFrame along with their corresponding transcripts.
  Entries without matching audio files were removed.
  The final cleaned data containing valid audio–text pairs was saved as metadata.csv in the working directory.
- **Preprocessing:**
  The generated metadata.csv file was used to inspect samples and verify correctness.
  A few random samples were selected to check the transcript and listen to the corresponding audio.
  This confirmed that the metadata and recordings were aligned and ready for fine-tuning.

2. Dataset 2

- **Location and Name:**
  Florence-2 OCR Testing Dataset Image Source:
  https://raw.githubusercontent.com/gokul-1998/handwriting_recognition/main/test_images/test3.png

- **Description:**
  A single test image containing handwritten or printed text was used to evaluate the OCR capabilities of the **Florence-2-Flux-Large** model.
  The image was loaded directly from a public GitHub repository URL and processed using the Florence-2 model architecture for text extraction.

- **Data Preparation:**
  The image was retrieved from the provided URL and opened in memory using Python's BytesIO.
  Before inference, the image was verified and converted to **RGB mode** to ensure compatibility with the model's input requirements.

- **Preprocessing:**
  The **Florence-2 AutoProcessor** was initialized with the model to handle both the text prompt and image input.
  A task prompt (<OCR>) was used to instruct the model to perform Optical Character Recognition.
  The processor tokenized the input text and prepared the image tensor for the model.
  The processed inputs were then passed to the model to generate the OCR output, which was decoded and post-processed to extract the recognized text.

3. Dataset 3

- **Name:** Automated Medical Diagnosis Using Clinical Notes - Kaggle
  - Location - https://www.kaggle.com/datasets/smmmmmmmmmmmm/automated-medical-diagnosis-using-clinical-notes
- **Description:** (from the data card) The synthetic dataset, containing 3000 simulated patient records, includes patient demographics, primary complaints, additional symptoms, and initial diagnostic suggestions. This dataset enables the LLM to learn patterns and correlations in clinical contexts, potentially assisting healthcare providers in decision-making and improving diagnostic efficiency in real-world medical environments.
- **Data Preparation:**
  - Single set of samples; training, val, and test splits to be done at model time
  - Reasonably clean data, No missing values in required features
  - EDA explained in GitHub - https://github.com/Pallavi-Pandey/DSAI_Lab_Project/blob/struct/notebooks/text/DSAIL_M2_part1.ipynb
  - Since it is synthetic data, it is well balanced in the comparisons done
- **Preprocessing:**
  - 3000 samples
  - May not need much preprocessing
  - Can be directly loaded to the model as required
- Notes: A larger dataset - MIMIC-IV-Ext-BHC - requires training and agreement signing; this could be included once these steps are completed