Task Assignment | AI Researcher Intern- Speech & Audio | Josh Talks

Question

Background
You are provided with ~10 hours of Hindi ASR training data (here) in the format shown below (audio + transcription metadata)

Dataset Schema Description

- **user_id** – Identifier for the speaker/user associated with the audio (anonymized).
- **recording_id** – Unique identifier for the specific audio recording within the dataset.
- **language** – Language label of the audio (e.g., "hi" for Hindi).
- **duration** – Duration of the audio recording (in seconds). Useful for filtering or batching.
- **rec_url_gcp** – URL link to the raw audio file stored on cloud (e.g., Google Cloud Storage). This is the main audio input for training/evaluation.
- **transcription_url** – URL to the ground-truth transcription text corresponding to the audio file. This is the label to be used for fine-tuning.
- **metadata_url** – URL to additional metadata about the recording (may include device type, noise level, accents, or collection conditions). Optional for training, but can help in analysis.

Your Task

a) Preprocess the dataset and share what you did to process the data and make it ready for training.
b) Fine-tune Whisper-small on this dataset and evaluate both the pretrained Whisper-small baseline and your fine-tuned model on the Hindi portion of the FLEURS test dataset.
c) Report the Word Error Rate (WER) in a structured table format. Here