To create a model monitoring pipeline for the audio speech recognition model, we will require a few things:

- Metrics
- Baseline performance
- Ground truth and predictions for new audio files
- Scheduled reporting and notifications

1. Metrics

- **Word Error Rate (WER):** Measures the sum of the number of words substituted, deleted, and inserted in the generated text, against the total number of words in the reference text. In the context of the database, WER is calculated using a global sum across all the records. The lower the value, the more accurate and better generated texts are.
- **Match Error Rate (MER):** Like the WER metric, instead of against the total number of words in the reference text, the sum is measured against the total number of words in both the reference and generated text

2. Baseline Performance

A test dataset can be used to measure the benchmark WER and MER for the audio speech recognition model that is currently deployed into production. We can make use of the audio files in the cv-valid-test folder found in the common-voice Kaggle dataset to calculate a baseline WER and MER. This baseline will be used to detect whether any model drift is within acceptable tolerance.

3. Ground truth and generated text for new files

For new audio files used in inference, the generated text can be consolidated saved in an object store like S3.

The ground truth for these new audio files can be obtained using manual transcription by an in-house, though it might be very resource intensive, or can be obtained from reference transcripts that are already available. Otherwise, sampling methods can be implemented to selectively perform manual transcription.

Both sets of data can be stored for future model evaluation and fine tuning of the model, and even form the new dataset to update the baseline performance

4. Scheduled reporting/notification

An orchestrator can trigger a serverless reporting pipeline (eg EventBridge, Lambda) on a scheduled basis (eg every weekend) to calculate the WER and MER of the new reference and generated texts. These metrics are measured against the existing baseline performance.

Custom rules can be set to detect acceptable or severe model drifts, such as the WER metric consecutively exceeding the acceptable threshold for 3 weeks. If severe model drift is detected (eg WER > 25%), a notification can be sent to the AI Engineering team to inform of the model drift, OR it can trigger a fine-tuning /deployment pipeline that uses the new data for fine-tuning.