8. Essay question – propose a model monitoring pipeline and describe how you would track model drift in 500 words.

Ground Truth Labels

This would depend on the use case. If we assume the new data to be indexed is unlabelled, for example audio that has been picked up from eavesdropping in a public space, then human volunteers could be recruited to transcribe a 10% sample of the data, i.e. create labels for accent, gender, and generated text.

Data Quality Metrics

Metrics could be gathered on the quality of the source data. For example, Signal to Noise Ratio and sampling rates could be checked for drift using the Wasserstein Distance. Missing metadata could be expressed as a percentage. In the common-voice dataset, most of the entries did not have accent/age/gender included.

Performance Metrics

The word-error-rate and character-error-rate are common performance metrics for ASR (word-error-rate is preferred for English). To quantify drift in these metrics, the data could be segmented into daily or monthly intervals, and the mean word-error-rate calculated for each batch. A time series graph could then be plotted to show how the batch word-error-rate changes over time. 'Drift' could be defined by setting a threshold for deviation as compared with a moving average of the past n batches. If the latest batch exceeds or drops below this threshold, then the model is deemed to have drifted.

The word-error-rate could be sliced along accent, gender, and age in order to characterise the model performance on different categories. The distribution of accents/gender/age could also be quantified using Frobenius Norm. These time series metrics and aggregated analyses could be fed to a Power BI or Tableau dashboard.

All these metrics should provide clues into what the source of drift in the underlying data is. If the sound quality is getting worse, a loss in performance could be expected across all categories. Or if an increasing proportion of audio is spoken in an accent that the model has difficulty with, a loss in overall performance would be expected. However if significant drift occurs within categories, then there may be inconsistency in the human labelling process, or some other confounding variable (e.g. a virus going round that gives old people a sore throat).

End User Monitoring

End users could also be recruited into the 'monitoring' process, by giving them a way of providing feedback. Each search result card could have a button that when clicked, reports the generated text as having an incorrect transcription. This feedback loop could save human transcribers time. In effect, they may only have to transcribe an initial 10% of the data for ground truth metrics. The other 90% would have a first pass done by the ASR model and only wrong entries would need to be corrected.

Finetuning Model Metrics

If the model underperforms on certain accents, it may be worth adding automatic accent identification as an initial stage, and routing those accents to a LoRA-enhanced version of the ASR model. Word-error-rates would have to be gathered for the training, validation, and test sets. A big divergence between training and test error rates may indicate overfitting.