Audio-to-Text Conversion - Evaluation layer

Core metrics (what/when/why)

- **WER (Word Error Rate)** = (S + D + I) / N *Use for:* most Latin-script languages and word-segmented scripts. *Insight:* overall transcript correctness; sensitive to word insertions/deletions.
- **CER (Character Error Rate)** *Use for:* languages without whitespace (zh, ja), noisy text normalization, or when tokenization is tricky. *Insight:* fine-grained errors; correlates with readability for non-segmented scripts.
- **Segment WER vs Concatenated WER** *Use for:* long-form audio evaluated by chunks. *Insight:* concatenated WER reveals stitch/overlap issues hidden by per-chunk scoring.
- **Entity/Number Accuracy (custom slots)** *Use for:* domains heavy in numerals, IDs, names (finance/medical). *Insight:* business-critical correctness beyond overall WER.
- **RTF (Real-Time Factor) & Latency (p50/p90)** *Use for:* streaming/production. *Insight:* deployment feasibility; RTF < 1 means faster-than-real-time offline decoding.
- **DER (Diarization Error Rate) & JER** *Use for:* multi-speaker ASR with speaker labels. *Insight:* speaker attribution quality (miss/false alarm/confusion).
- **LID Accuracy / Code-switch WER** *Use for:* multilingual pipelines. *Insight:* language routing quality; per-language WER comparisons.
- **Calibration (ECE/Brier/NLL on confidence)** *Use for:* post-ASR confidence scoring. *Insight:* how well scores reflect true correctness (useful for human-in-the-loop).

> Practical rule: **WER/CER** for core model progress, **DER** when speakers matter, **entity/number accuracy** for domain usefulness, **RTF/latency** for deployability.

Visualization methods (to diagnose + explain)

- **Alignment heatmaps**
  - *Seq2seq attention maps:* decoder-to-encoder attention over time (token ↔ frame).
  - *CTC alignments:* frame-level best path / forced alignment overlay on spectrogram.
- **Word-timeline plots**
  - Show predicted words with start/end times over a waveform or log-Mel spectrogram.
- **Error overlays**
  - Color Levenshtein operations (S/D/I) along the timeline; spotlight where/why WER arises.
- **Entity highlighting**

  - Highlight numbers/dates/tickers in reference vs hypothesis (correct/incorrect).
- **Diarization ribbons**
  - Horizontal bars per speaker with ASR text above; quickly reveals overlap/confusions.
- **Saliency on spectrogram (Integrated Gradients)**
  - Attribute which time–freq regions influenced a token; great for noise/debugging.
