Assignment Submission Report (Questions 1 to 4)

Question 1

Objective

Preprocess ~10 hours Hindi ASR data, fine-tune Whisper-small, and compare baseline vs fine-tuned performance.

Methodology (Short)

* Filtered metadata for Hindi records and resolved broken URLs to upload_ goai format.

* Downloaded audio + transcription JSON and created segment-level training clips.

* Built train/ validation Hugging Face dataset for Whisper.

* Fine-tuning run started but stopped due to GPU limits (partial run only).

Preprocessing Summary

* Input Hindi recordings: 104

* Prepared segment samples: 5794

* Train rows: 5214

* Validation rows: 580

* Failures: 0

Training/ Eval Status (Partial)

* Training progress at interruption: [401/ 800], Epoch 1.23/ 3

* Baseline WER (reported): 0.830 (83.00%)

* Validation WER at step 200: 0.418857 (41.89%)

* Validation CER at step 200: 0.209089 (20.91%)

* Relative WER improvement vs baseline: 49.54%

* Note: final fine-tuned FLEURS evaluation pending full training completion.

Q1 Deliverables

* Consolidated report section: outputs/ final_ report/ final_ report_ q1_ q4.md

* Q1 structured report: outputs/ q1_ report.md

* Q1 WER table: outputs/ q1_ wer_ table.csv

Question 2

Objective

Detect target Hindi speech disfluencies from segment transcripts, extract corresponding audio clips, and create a structured occurrence-level sheet.

Methodology (Short)

* Loaded Hindi metadata and resolved corrected upload_ goai URLs for transcript/ audio.

* Used hybrid disfluency detection:

* Lexicon matching from provided disfluency list.

* Regex rules for repetition, prolongation, and false-start patterns.

* For each segment with a detected disfluency, clipped audio from full recording using segment start/ end timestamps.

* Saved one row per disfluency occurrence in CSV with clip path and metadata.

Output Summary

* Input recordings processed: 104

* Disfluency occurrences detected: 7926

* Recordings with at least one hit: 104

* Failures: 0

Deliverables

* Sheet (occurrence-level): outputs/ q2_ disfluency_ segments.csv

* Segmented clips directory: data/ q2_ disfluency_ clips

* Summary JSON: outputs/ q2_ disfluency_ summary.json

* Methodology note: outputs/ q2_ methodology.md

Question 3

Objective

Classify unique words into:

* correct spelling

* incorrect spelling

Methodology (Short)

* Used the unique-word file (Unique Words Data - Sheet1.csv).

* Applied Unicode normalization and Hindi orthography checks:

* Non-Devanagari/ script-noise detection.

* Invalid sequence/ sign rules (obvious spelling/ character errors).

* Produced two-column output: word, spelling_ label.

Output Summary

* Total unique words processed: 175,780

* Correct spelling: 148,396

* Incorrect spelling: 27,384

Deliverables

* Output sheet: outputs/ q3_ word_ spelling_ labels.csv

* Summary JSON: outputs/ q3_ spelling_ summary.json

* Methodology note: outputs/ q3_ methodology.md

Question 4

Objective

Evaluate model transcripts against human reference transcripts and select best-performing model using WER/ CER.

Methodology (Short)

* Used Human as reference in Question 4 - Task.csv.

* Normalized text and computed aggregate WER/ CER for:

* Model H, Model i, Model k, Model l, Model m, Model n

* Ranked models by WER (primary) then CER.

* Generated per-segment analysis for deeper review.

Output Summary

* Input segments: 46

* Best model: Model i

* Best WER: 0.001222

* Best CER: 0.000829

Deliverables

* Model metrics table: outputs/ q4_ model_ metrics.csv

* Segment-level analysis: outputs/ q4_ segment_ analysis.csv

* Summary JSON: outputs/ q4_ summary.json

* Short report: outputs/ q4_ report.md

Reproducibility (Scripts Used)

* Q1: scripts/ prepare_ joshtalk_ hindi.py, scripts/ train_ whisper_ hindi.py, scripts/ evaluate_ fleurs_ hi.py

* Q2: scripts/ prepare_ disfluency_ dataset.py

* Q3: scripts/ classify_ q3_ spelling.py

* Q4: scripts/ evaluate_ q4_ models.py