CORAL

Consensus-based Refinement And Learning

A Research & Development Final Year Project

Presenters:

- Muhammad Rafay Khan Khattak (21I-0423)
- Muhammad Ali Irfan (21I-2572)
- Muhammad Nouman Hafeez (21I-0416)

Supervisor:

- Ms. Kainat Iqbal

Co-supervisor:

- Ms. Saira Qamar

8/29/2025

CORAL Overview

Turn multiple ASR outputs into one more-accurate Urdu transcript

Designed to work with existing ASR models without retraining.

If models disagree, CORAL picks the word that looks most reliable while keeping the sentence natural.

Outputs per-word confidence scores so other applications can decide which parts to trust.

Motivation

Live captions for TV, lectures, and online classes

Call-center & customer support analytics

Voice assistants that understand code-switched Urdu-English

Clearer subtitles for deaf and non-native viewers

Problem Statement

“Single-model Urdu ASR systems fail to reliably resolve acoustic and linguistic ambiguity, leading to high WER. We need a confidence-guided ensemble correction method”

LITERATURE REVIEW

Paper	Reference	Year	Approach	Key features	Deficiencies
Better Pseudo-labeling with Multi-ASR Fusion and Error Correction by SpeechLLM	https://arxiv.org/abs/2506.11089	2025	Multi-ASR ensemble (Icefall, Nemo, Whisper) fused and corrected by an instruction-tuned SpeechLLM.	**Unified pipeline using an LLM to refine ASR hypotheses with both audio and text cues, yielding near ground-truth labels.**	**High computational cost and latency. . Requires multiple pre-trained ASR models for the target language.**
ASR Confidence Estimation using True Class Lexical Similarity Score (TruCLeS)	https://www.isca-archive.org/interspeech_2025/ravi25_interspeech.pdf	2025	Trains a neural confidence model using a novel, continuous word-level confidence target (TruCLeS).	**Provides a fine-grained confidence score reflecting partial correctness. Model-agnostic and tested on Hindi, showing generality.**	**Requires forced alignment for training. Does not directly reduce WER. Does not address using confidences for correction.**
Leveraging LLMs for Post-Transcription Correction	https://arxiv.org/pdf/2506.11089	2024	Post-ASR correction using a retrieval-augmented LLM (GPT-3.5) to find and replace domain-specific errors.	**Model-agnostic and effective for fixing critical, known terms in a specific domain by leveraging LLM knowledge.**	**Highly specialized to known target words. Requires access to powerful LLMs and historical data. English business data only.**
Ensembles of Hybrid and End-to-End Speech Recognition	https://aclanthology.org/2024.lrec-main.547.pdf	2024	Combines a hybrid HMM-Kaldi ASR with a wav2vec2.0 XLS-R model using confidence-calibrated ROVER.	**Achieves significant WER reduction (14-20%) by combining complementary strengths and addresses E2E model overconfidence.**	**Requires training two separate ASR systems. Tested only on a low-resource European language. ROVER is sensitive to alignment errors.**
Code-Mixed Street Address Recognition	https://www.researchgate.net/publication/385763612_Code-mixed_street_address_recognition_and_accent_adaptation_for_voice-activated_navigation_services	2024	Builds a custom hybrid TDNN-LSTM ASR system trained on bespoke Urdu-English datasets for street addresses.	**Achieves very low WER (~4.0%) in a narrow domain. Explicitly handles code-mixing and accent adaptation for Urdu.**	**Does not generalize outside its specific application. Relies on cumbersome hybrid architecture and custom data.**

Challenges

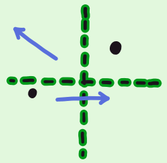
Match words despite splits or missing parts

Confidence can be wrong

Latency and cost

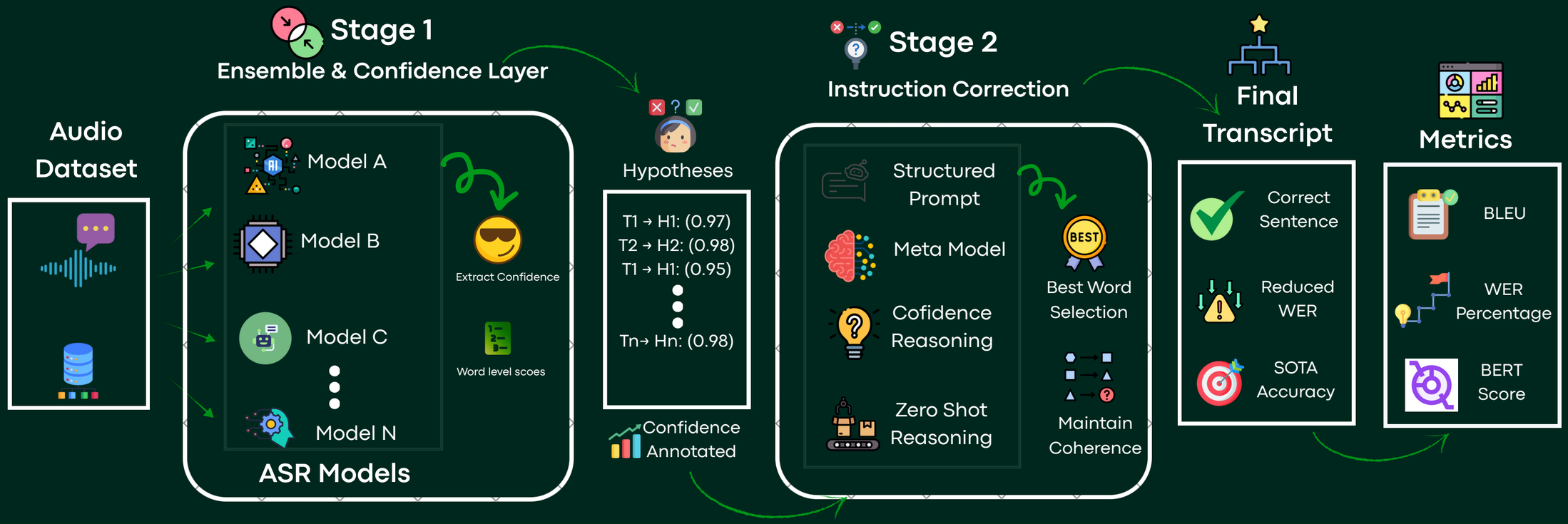
Models give confidence differently

LLMs may hallucinate



Test across languages

Proposed Solution



1. Multi-ASR Decoding

Run multiple ASR models in parallel to get word tokens, timestamps, and confidence scores for every token.

2. Confidence Normalization

Convert raw model scores to a uniform 0–1 scale and calibrate them using held-out data.

3. LLM Fusion

Provide aligned alternatives with their confidences to an instruction-tuned LLM. The LLM outputs a single transcript.

4. Comparison with Baselines

Evaluate against ROVER (voting), confusion networks, and highest-confidence fusion.

TIMELINE

Iteration 1

(Sep - Oct 2025)

Foundation & Confidence Extraction

- Integrate ASR models
- Extract word-level confidences
- Establish baselines.

Iteration 2

(Nov - Dec 2025)

Instruction Prompt Development

- Design prompts
- Test LLM behavior
- Structured input/output parsing

Iteration 3

(Feb - Mar 2026)

End-to-End Integration

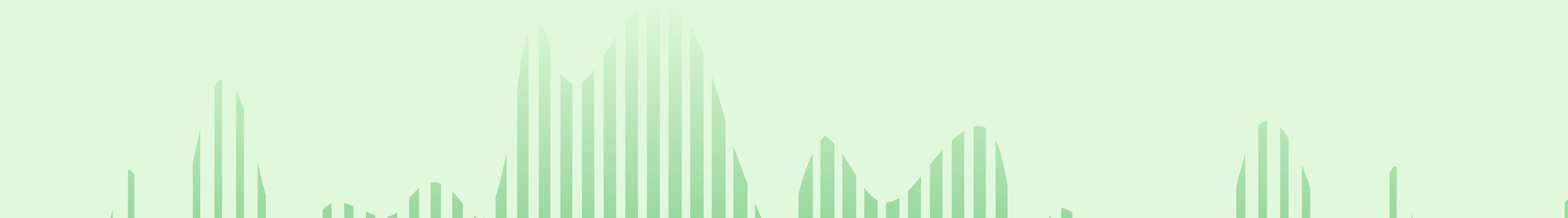
- Integrate pipeline
- Run preliminary evaluations

Iteration 4

(Apr - May 2026)

Optimization & Comprehensive Evaluation

- Calibrate
- Profile latency, optimize
- Prepare final report



WORK DIVISION

- ASR ensemble integration
- Confidence extraction
- Benchmarking
- N-best & oracle evaluation

- Black-box LLM integration
- Prompt engineering
- Alignment parser for LLM input
- Final fusion evaluation

- System architecture
- Optimization (quantization/ONNX)
- Deployment-ready pipeline and latency profiling

CORAL Project Timeline - Gantt Chart

Tools and Technologies

Python

Jupyter notebook

Scikit-learn

Pytorch

Huggingface

GitHub

Docker

Runpod

React

Conclusion

- CORAL uses word-level confidence + LLMs for zero-shot Urdu ASR refinement.
- No fine-tuning; combines pre-trained models for better generalization.
- Finalize experiments, cut LLM latency, deploy demo, prepare results.

REFERENCES

- Prakash, J., et al. (2025). Better Pseudo-labeling with Multi-ASR Fusion and Error Correction by SpeechLLM. In Proceedings of Interspeech 2025.
- Nagarathna R., et al. (2025). ASR Confidence Estimation using True Class Lexical Similarity Score (TruCLeS). In Proceedings of Interspeech 2025.
- Koilakuntla, B., et al. (2024). Leveraging Large Language Models for Post-Transcription Correction in Contact Centers. In Proceedings of Interspeech 2024.
- Parikh, A. K., et al. (2024). Ensembles of Hybrid and End-to-End Speech Recognition. In Proceedings of LREC-COLING 2024.
- Naqvi, S. M. R. & Tahir, M. A. (2024). Code-Mixed Street Address Recognition and Accent Adaptation for Voice-Activated Navigation Services. IEEE Access.
- Radford, A., et al. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356.
- Mohiuddin, H. A., et al. (2023). UrduSpeakXLSR: Multilingual Model for Urdu Speech Recognition. In 2023 18th International Conference on Emerging Technologies (ICET).
- Kam, E. M. W., & Uebel, L. (2021). On Confidence Score Generation for Transformer-based ASR. In Proceedings of Interspeech 2021.

THANK YOU

FROM TEAM CORAL