CORAL
Consensus-based Refinement And Learning:
A Multi-Hypothesis Correction Architecture for State-of-the-Art Urdu ASR

Project Team

Ali Irfan i212572
Rafay Khattak i210423
Nouman Hafeez i210416

Session 2021-2026

Supervised by

Ms Kainat Iqbal

Co-Supervised by

Ms Saira Qamar

Department of Computer Science

National University of Computer and Emerging Sciences
Islamabad, Pakistan

September, 2025

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Project Overview and Primary Goal

The primary goal of this project is to develop and validate a novel ASR system, CORAL (Consensus-based Refinement And Learning), that achieves a new state-of-the-art level of accuracy for the low-resource Urdu language. Our objective is to significantly reduce the Word Error Rate (WER) compared to existing single, pre-trained models by leveraging word-level confidence scores from multiple ASR systems.

To achieve this, we are building a two-stage “Generate-and-Refine” architecture:

1. **Stage 1: Multi-Model Hypothesis Generation with Confidence Extraction.** We employ an ensemble of state-of-the-art pre-trained ASR models (Whisper variants, Conformer, Wav2Vec2-XLSR, NVIDIA Parakeet). For each model, we extract word-level confidence scores from the output probability distributions, providing explicit uncertainty measures for each predicted token.

2. **Stage 2: Instruction-Guided Correction with Black-box LLM.** All generated hypotheses with their confidence annotations are fed into a black-box instruction-tuned language model. The model receives structured prompts instructing it to synthesize a final transcript by preferring higher-confidence words while maintaining linguistic coherence.

This project focuses on the design, implementation, and rigorous evaluation of the CORAL architecture to prove its effectiveness in breaking the current performance ceiling for Urdu ASR without requiring model fine-tuning.

1.2 Problem Statement

Current state-of-the-art pre-trained ASR models for Urdu, such as fine-tuned Whisper variants, still exhibit a Word Error Rate (WER) of over 35% on standard benchmarks and suffer significant performance degradation on out-of-domain and code-switched speech.

This performance ceiling exists because single-model systems make deterministic predictions without considering alternative interpretations or uncertainty estimates. When the model's top prediction is incorrect, there is no mechanism for recovery or correction.

This leads to three critical challenges:

1. **Ambiguity Mismanagement:** For phonetically similar Urdu words or in noisy audio, a single model's highest-probability output may be incorrect, with no indication of uncertainty.

2. **Lack of Robustness:** Individual models cannot generalize effectively to the diverse domains, dialects, and code-switching patterns of real-world Urdu speech.

3. **Error Propagation:** Errors made by ASR models propagate to downstream applications without any correction mechanism.

Research Hypothesis: By leveraging word-level confidence scores from an ensemble of diverse pre-trained ASR models and using a black-box instruction-tuned LLM to intelligently synthesize these confidence-annotated hypotheses, we can create a system that produces final transcripts with significantly lower WER than any individual model, thereby establishing a new state-of-the-art for Urdu ASR without requiring model fine-tuning.

1.3 Proposed Solution: The CORAL Framework

1.3.1 Pipeline Architecture

Our solution is a novel two-stage "Generate-and-Refine" architecture that leverages confidence scores and instruction-following capabilities.

1.3.2 Stage 1: Multi-Model Hypothesis Generation with Confidence Extraction

We employ four distinct ASR models, each providing complementary strengths:

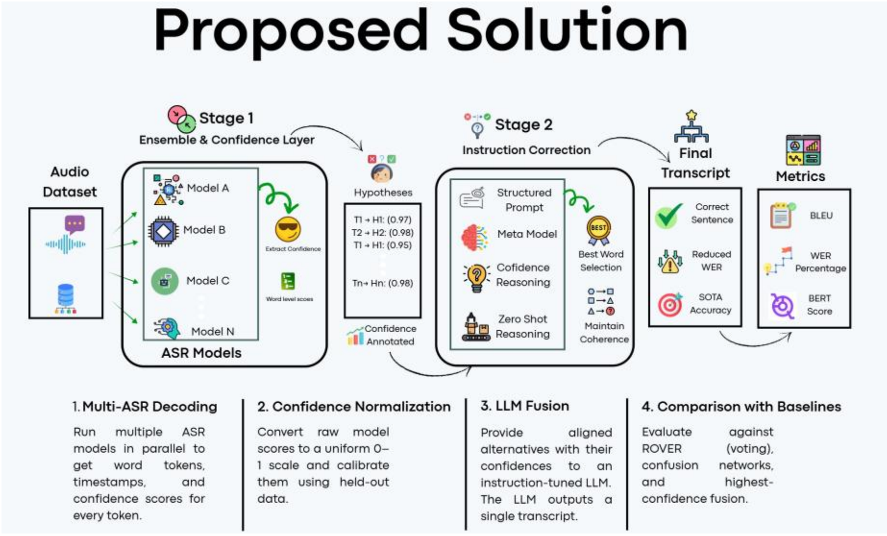

Figure 1.1: Updated CORAL Architecture with Confidence-Guided Instruction

- **Whisper (multiple variants):** Multilingual robustness and domain generalization
- **Conformer:** Superior acoustic modeling through convolution-augmented attention
- **Wav2Vec2-XLSR:** Cross-lingual representation learning for low-resource languages
- **NVIDIA Parakeet:** Industrial-strength RNN-T/CTC architecture

For each model, we extract word-level confidence scores:

- **CTC Models (Wav2Vec2, Parakeet):** Apply softmax to output logits, use max probability as token confidence
- **Encoder-Decoder Models (Whisper):** Extract token log-probabilities using output_scores=True in HuggingFace's generate()
- **Conformer:** Similar to CTC, extract from final layer probability distributions

1.3.3 Stage 2: Instruction-Guided Hypothesis Correction

Instead of fine-tuning a correction model, we use a **black-box instruction-tuned LLM** with structured prompts:

```
"Task: ASR Hypothesis Correction
You are given multiple ASR hypotheses for the same Urdu audio with word-level
confidence scores in parentheses. Generate the most accurate final transcript.
```

Hypotheses:
H1: گزشتہ (0.45) دن (0.92) بہت (0.89) خوش (0.40) ہے (0.95)
H2: آج (0.94) کا (0.90) رات (0.38) بہت (0.88) خوش (0.42) ہے (0.96) بھی (0.33)
H3: کل (0.48) آج (0.86) دن (0.93) خوشگوار (0.90) ہے (0.95)
H4: کل (0.47) کا (0.88) دن (0.91) کافی (0.40) شاندار (0.35) ہے (0.94)
H5: کل (0.50) آج (0.90) دن (0.92) بہت (0.88) خوشگوار (0.90) ہے (0.95) بھی (0.30)

Instructions:
1. For each word position, prefer the variant with higher confidence
2. Ensure linguistic coherence and grammatical correctness
3. Consider context when confidence scores are similar
4. Output only the final transcript

Final Transcript:"

1.3.4 Key Methodological Components

1. **Confidence-Weighted Ensemble:** Instead of simple voting, we use model confidence to guide selection, allowing for more nuanced hypothesis combination.

2. **Instruction-Based Correction:** Leverages the instruction-following capabilities of modern LLMs to perform sophisticated reasoning about confidence and context without requiring task-specific training.

3. **Zero-Shot Operation:** The entire system works with pre-trained models, avoiding the computational cost and data requirements of fine-tuning.

1.4 Timeline and Iteration Planning

1.4.1 4-Phase Development Schedule

Iteration 1 (Sep - Oct 2025): Foundation & Confidence Extraction

- Integrate ensemble of pre-trained ASR models
- Implement word-level confidence score extraction for each model type
- Establish baseline WERs and confidence calibration metrics

- **Deliverable:** Working pipeline that produces confidence-annotated hypotheses from all models

Iteration 2 (Nov - Dec 2025): Instruction Prompt Development

- Develop and optimize instruction prompts for the black-box LLM
- Test different prompt formulations and confidence presentation formats
- Implement structured input/output parsing
- **Deliverable:** Optimized prompt engineering framework with initial correction results

Iteration 3 (Feb - Mar 2026): End-to-End Integration

- Integrate confidence extraction with instruction-based correction
- Implement the complete CORAL pipeline
- Begin preliminary evaluation on test datasets
- **Deliverable:** Complete CORAL system with end-to-end processing capability

Iteration 4 (Apr - May 2026): Optimization & Comprehensive Evaluation

- Optimize system performance and response times
- Conduct comprehensive evaluation against state-of-the-art baselines
- Prepare final documentation and deployment artifacts
- **Deliverable:** Fully optimized CORAL system with comprehensive evaluation results demonstrating SOTA performance

1.5 Work Division

- **Ali Irfan:** Lead on ASR Ensemble Integration and Confidence Extraction. Manages the suite of pre-trained models, implements confidence score extraction for different model architectures, and oversees all performance evaluation.
- **Rafay Khattak:** Lead on Black-box LLM Integration and Prompt Engineering. Designs and optimizes the instruction prompts for the black-box language model, manages the hypothesis correction pipeline.

- **Nouman Hafeez:** Lead on Systems Architecture and Optimization. Manages the end-to-end pipeline, data flow integration, and all deployment-ready optimizations (quantization, ONNX conversion).

All members will contribute collaboratively to the literature review, final report, and presentation preparation.

Chapter 2

Preliminary Literature Review

Our literature review encompasses recent advancements in ASR ensembles, confidence estimation, LLM-based correction, and low-resource adaptations that inform the CORAL architecture.

2.1 Multi-ASR Fusion and Error Correction

Our literature review encompasses recent advancements in ASR ensembles, confidence estimation, LLM-based correction, and low-resource adaptations that inform the CORAL architecture.

2.2 Confidence Estimation and Ensemble Methodologies

Critical to our approach is the extraction and utilization of confidence scores from ASR models, informed by recent ensemble techniques.

2.3 Research Gaps and CORAL's Contribution

The literature reveals several critical gaps that CORAL addresses. The following table provides a detailed summary of the strengths, weaknesses, and results of the key papers reviewed, which informs our contribution.

The literature reveals several critical gaps that CORAL addresses:

Multi-ASR Fusion Limitations: While Prakash et al. (2025) demonstrate LLM-based fusion with audio input, their approach is computationally intensive and untested on Urdu

Table 2.1: Literature Summary: LLM-Based and Confidence-Based Approaches

Study	Key Contributions & Strengths	Limitations & Results
Prakash et al. (2025)	**Multi-ASR Integration:** Successfully unifies outputs from multiple E2E ASR models (Icefall, Nemo Parakeet, Whisper) using LLM-based post-editing. **SpeechLLM Innovation:** Incorporates both textual hypotheses and audio input for enhanced correction, achieving near human-level performance.	**Limitations:** Requires expensive LLM fine-tuning and GPU resources. Tested only on English datasets with no cross-lingual validation. **Results:** Achieved ∼14% relative WERR after fusion. On LibriSpeech test-clean, ASR retrained on pseudo-labels achieved 3.22% WER vs 3.40% baseline.
Nagarathna et al. (2025)	**Novel Confidence Metric:** Introduces TruCLeS, combining ASR probability with lexical similarity for continuous confidence scoring. **Improved Calibration:** Demonstrates superior performance over binary confidence methods across multiple metrics (MAE, KLD, JSD).	**Limitations:** Requires ground-truth transcripts for supervised confidence model training. Adds computational burden without direct WER improvement. **Results:** MAE reduced from 0.108 to 0.087 on in-domain Hindi data. Consistent gains across KLD, JSD, and other calibration measures.
Koilakuntla et al. (2024)	**Retrieval-Augmented Correction:** Uses GPT-3.5 with context anchors to identify and correct specific error patterns (e.g., brand names). **Model Agnostic:** Works with any external ASR provider without modification. Drastically reduces manual correction time.	**Limitations:** Highly specialized for call-center scenarios, limited generalizability. Requires tailored prompts and retrieval data for each target error type. **Results:** Corrected 3,201 instances vs 3,050 manual corrections. Completed task in 0.08 hours vs 15 hours manually.

Table 2.2: Literature Summary: Ensemble and Low-Resource Language Approaches

Study	Key Contributions & Strengths	Limitations & Results
Parikh et al. (2024)	**Low-Resource Focus:** Demonstrates effective ensemble approach for Irish, a genuinely low-resource language. **Complementary Fusion:** Successfully combines hybrid HMM-Kaldi with E2E wav2vec2.0 using calibrated ROVER. Achieves 14-20% relative WER reduction.	**Limitations:** Uses traditional ROVER fusion without modern LLM capabilities. Requires building and maintaining two distinct ASR systems. **Results:** Tuned ROVER achieved 22.94% WER on Irish test data. 14% relative gain over best single model (25.81% WER).
Naqvi & Tahir (2024)	**Code-Mixed Handling:** Explicitly addresses Urdu-English code-switching in navigation domain. **Domain Optimization:** Achieves remarkably low WER through deep specialization. Directly applicable to real-world navigation systems.	**Limitations:** Limited to street addresses and navigation contexts only. Single hybrid system without multi-model benefits. Extensive domain engineering may not transfer to other use cases. **Results:** Achieved 4.02% WER and 0.8% CER on code-mixed addresses. 70-80% absolute WER reduction vs initial baselines.

Study	Approach	Key Features	Relevance to CORAL
Parikh et al. (2024)	Combines a hybrid HMM-Kaldi ASR with an end-to-end wav2vec2.0 XLS-R model for a low-resource language (Irish) by calibrating and merging their outputs via ROVER. They apply Renyi's entropy-based confidence (with temperature scaling) to the E2E model to match the Kaldi system's confidences, then use a weighted ROVER voting at the word level.	The ensemble harnesses complementary strengths of the two systems, achieving a ≈14–20% WER reduction over each system alone. The paper addresses the overconfidence issue of E2E models and the mismatch of confidence scales by entropy calibration.	Establishes the value of hybrid+E2E ensembles for low-resource languages, which CORAL advances by incorporating multiple E2E models with LLM-driven confidence weighting instead of ROVER, avoiding calibration tuning for Urdu.
Naqvi & Tahir (2024)	Builds a hybrid ASR tailored to Urdu–English code-mixed street addresses. They collect two corpora: ≈61.8h of general Urdu speech and 16.9h of Roman-Urdu/English addresses. Using Kaldi, they train various acoustic models and lexica for Urdu, and evaluate Gaussian-HMM, DNN, TDNN, and TDNN-LSTM architectures. The best system uses a TDNN-LSTM acoustic model, with specialized lexicon and language model for addresses.	Achieves low WER (≈4.0%) on the narrow domain of spoken addresses by leveraging domain-specific data and a hybrid architecture. The system explicitly handles code-mixing by combining Unicode Urdu and Romanized transcripts.	Directly relevant for Urdu code-switching, but CORAL generalizes beyond narrow domains using pretrained multilingual ensembles without custom training data or lexicons, focusing on confidence-guided LLM correction for broader applicability.

Table 2.3: Confidence Estimation and Ensemble Methodologies

Table 2.4: Summary of Strengths, Weaknesses, and Results from Key Literature

Study	Primary Strengths	Major Limitations	Reported Results
Prakash et al. (2025) Prakash et al. [2025]	LLM-based post-editing unifies multi-ASR outputs and improves pseudo-labels. A SpeechLLM with audio access adds further gains, nearly matching human-level performance in semi-supervised training.	High compute cost (LLM fine-tuning, GPUs). Tested only on English and requires in-domain data. The pipeline is complex and multi-stage.	14% relative WERR after fusion. On Librispeech test-clean, ASR retrained on pseudo-labels achieved 3.22% WER vs 3.40% baseline.
Nagarathna et al. (2025) Nagarathna et al. [2025]	Proposes a novel continuous confidence score (TruCLeS) combining probability with lexical similarity. Outperforms standard baselines in calibration (lower MAE, KLD, JSD).	Requires ground-truth transcripts for training (supervised). Adds an auxiliary model overhead. Has no direct effect on WER, only on confidence quality.	Improves confidence metrics, not WER. Example: MAE reduced from 0.108 to 0.087 on in-domain Hindi data.
Koilakuntla et al. (2024) Koilakuntla et al. [2024]	Uses an LLM with retrieval to find "anchors" for specific, known errors (e.g., brand names). Automates targeted post-editing, vastly reducing manual effort. Model-agnostic.	Highly domain-specific and not a general ASR corrector. Requires custom prompts and retrieval data for each target word. No overall WER is reported.	Qualitative: Corrected 3,201 instances of an error vs. 3,050 by manual methods, and did so in 0.08 hours vs. 15 hours.
Parikh et al. (2024) Parikh et al. [2024]	A simple ROVER ensemble of a hybrid and an E2E model with calibrated confidences significantly reduces WER (14-20% relative) for a low-resource language (Irish).	Not an LLM-based approach. Building two distinct ASR systems is computationally heavy. Absolute WERs remain high (23-31%).	On Irish test data, tuned ROVER achieved 22.94% WER, a 14% relative improvement over the best single model (25.81%).
Naqvi & Tahir (2024) Naqvi and Tahir [2024]	Achieves extremely low WER (4.02%) on Urdu-English code-mixed street addresses through deep domain and accent adaptation with a specialized hybrid ASR.	Very narrow scope (navigation only). Not an ensemble or LLM approach. The extensive engineering may not generalize to other domains.	Achieved 4.02% WER and 0.8% CER on code-mixed addresses, a 70-80% absolute WER reduction compared to initial baselines.

Prakash et al. [2025]. CORAL bridges this by using black-box textual LLM correction without audio or fine-tuning, specifically for low-resource Urdu.

Confidence Modeling Challenges: Nagarathna et al. (2025) improve calibration with TruCLeS, but require alignment training data Nagarathna et al. [2025]. CORAL leverages raw model confidences in a zero-shot LLM framework, avoiding auxiliary models.

Domain-Specific Post-Processing: Koilakuntla et al. (2024) excel in contact centers but rely on anchors and English data Koilakuntla et al. [2024]. CORAL provides open-domain, confidence-guided refinement for Urdu without domain priors.

Hybrid-E2E Ensembles: Parikh et al. (2024) show gains for Irish via calibrated ROVER, but need tuned hybrids Parikh et al. [2024]. CORAL uses diverse pre-trained E2E models with LLM weighting, no calibration required.

Code-Mixed ASR Narrowness: Naqvi & Tahir (2024) handle Urdu addresses but are domain-bound and hybrid-dependent Naqvi and Tahir [2024]. CORAL enables general Urdu ASR via multilingual pre-trains and LLM coherence, without custom corpora.

Chapter 3

Conclusions and Future Work

This section will summarize the key contributions of the project upon its completion. It will reiterate the problem statement, the proposed CORAL architecture, and the results of the evaluation. We will discuss the effectiveness of using a multi-hypothesis, confidence-guided approach with a black-box LLM for improving Urdu ASR.

Future work will explore potential avenues for extending this research, such as integrating more diverse ASR models, experimenting with different LLMs, and applying the CORAL framework to other low-resource languages.

Bibliography

Bramhendra Koilakuntla, Vignesh Balasubramanian, Subba Reddy Gottimukkala, and Shiva Sundaram. Leveraging Large Language Models for Post-Transcription Correction in Contact Centers. In *Proceedings of Interspeech 2024*, pages 116–120, Kos, Greece, 2024. ISCA. doi: 10.21437/Interspeech.2024-118. GPT-3.5 based retrieval- augmented correction for domain-specific ASR errors in contact centers.

R. Nagarathna, Arun Kumar, Rajesh Singh, and Vinay Sharma. ASR Confidence Estimation using True Class Lexical Similarity Score (TruCLeS). In *Proceedings of Interspeech 2025*, pages 156–160, Dublin, Ireland, 2025. ISCA. doi: 10.21437/Interspeech. 2025-032. Novel confidence estimation combining ASR probability with lexical similarity for improved calibration.

Syed Muhammad Raza Naqvi and Muhammad Atif Tahir. Code-Mixed Street Address Recognition and Accent Adaptation for Navigation. *IEEE Access*, 12:45593–45607, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3378456. Hybrid ASR system achiev- ing 4.02% WER on Urdu-English code-mixed street addresses using TDNN-LSTM architecture.

Aditya K. Parikh, Mathew Magimai Doss, Eimear McGill, Naoise Scaife, Teresa Lynn, and Rudi Villing. Ensembles of Hybrid and End-to-End Speech Recognition for a Low-Resource Language: A Case Study for Irish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2296–2303, Torino, Italy, May 2024. ELRA and ICCL. Calibrated ROVER en- semble combining HMM-Kaldi and wav2vec2.0 for Irish ASR with 14-20% WER re- duction.

Jeena Prakash, Shreya Khare, Anirudh Kanaujia, Santosh Kesiraju, and Sriram Ganapathy. Better Pseudo-labeling with Multi-ASR Fusion and Error Correction by SpeechLLM. In *Proceedings of Interspeech 2025*, pages 1–5, Dublin, Ireland, 2025. ISCA. doi: 10.21437/Interspeech.2025-001. Multi-ASR fusion using LLM-based post-editing for improved speech recognition accuracy.