Automatic Speech Recognition for Telugu

Problem Statement

- Telugu is a widely spoken but under-resourced language in ASR.
- Most ASR systems are optimized for English or high-resource languages.
- Lack of accessible voice tech in Telugu limits inclusivity in digital applications.

Motivation and Importance

- Spoken by over 95 million people.
- Enables digital inclusion and accessibility.
- Useful in education, customer service, healthcare, and more.

Applications

- Documentation of local history: preserving the rich oral history, folk stories, regional customs, and cultural practices that are passed down verbally across generations
- Speech-to-Text for Hearing Impaired: Converts spoken Telugu into text to assist individuals with hearing impairments.
- Voice-Controlled Assistants: Enables voice-based interactions for visually impaired users.
- Real-Time Subtitling: Helps generate live captions for videos, meetings, news broadcasts, Movies and TV Shows
- Voice-Based Grievance Redressal: Helps citizens file complaints and queries in Telugu through speech.
- E-Governance Portals: Enables voice-driven interaction for public services, making them accessible to non-literate users.

User Stories

1. Rural Citizen Using Government Services
As a farmer who only speaks Telugu, I want to speak into a mobile app and get my message transcribed so I can file complaints or request government services without needing to write in English.

2. Healthcare Worker in a Village
As a health volunteer in a remote village, I want to use voice-to-text in Telugu so I can quickly record patient data without needing internet or typing tools.

3. NGO Field Worker
As a field worker for an NGO, I want to record interviews and community feedback in Telugu and automatically get clean text transcriptions for reporting and documentation.

User Stories

4. Content Creator
As a Telugu YouTuber, I want to transcribe my audio into accurate subtitles automatically so that I can reach a wider audience and improve accessibility.

5. Visually Impaired User
As a visually impaired person, I want to speak into my phone and have it transcribe my speech into text, so I can send messages, fill forms, or write notes independently.

6. Hearing Impaired User
As a person with hearing impairment, I want to use the app to convert spoken Telugu from others into text in real time, so I can follow conversations without relying on a sign language interpreter.

Why Whisper

- Whisper is trained on 680,000+ hours of multilingual and multitask data.
- Robust to Real-World Conditions
- Handles noisy data well

Architecture Diagram

Fine-Tuning of Transformer Based ASR model

The dataset includes speech from Open SLR, Indic TTS and Common Voice

Indic knowledge base

Load and split dataset

Selection of Audio and Transcription columns

Load Whisper Processor

The required sampling rate is 16kHz. Apply preprocessing to dataset and filter out audio clips exceeding 30 seconds

Resample audio to required rate

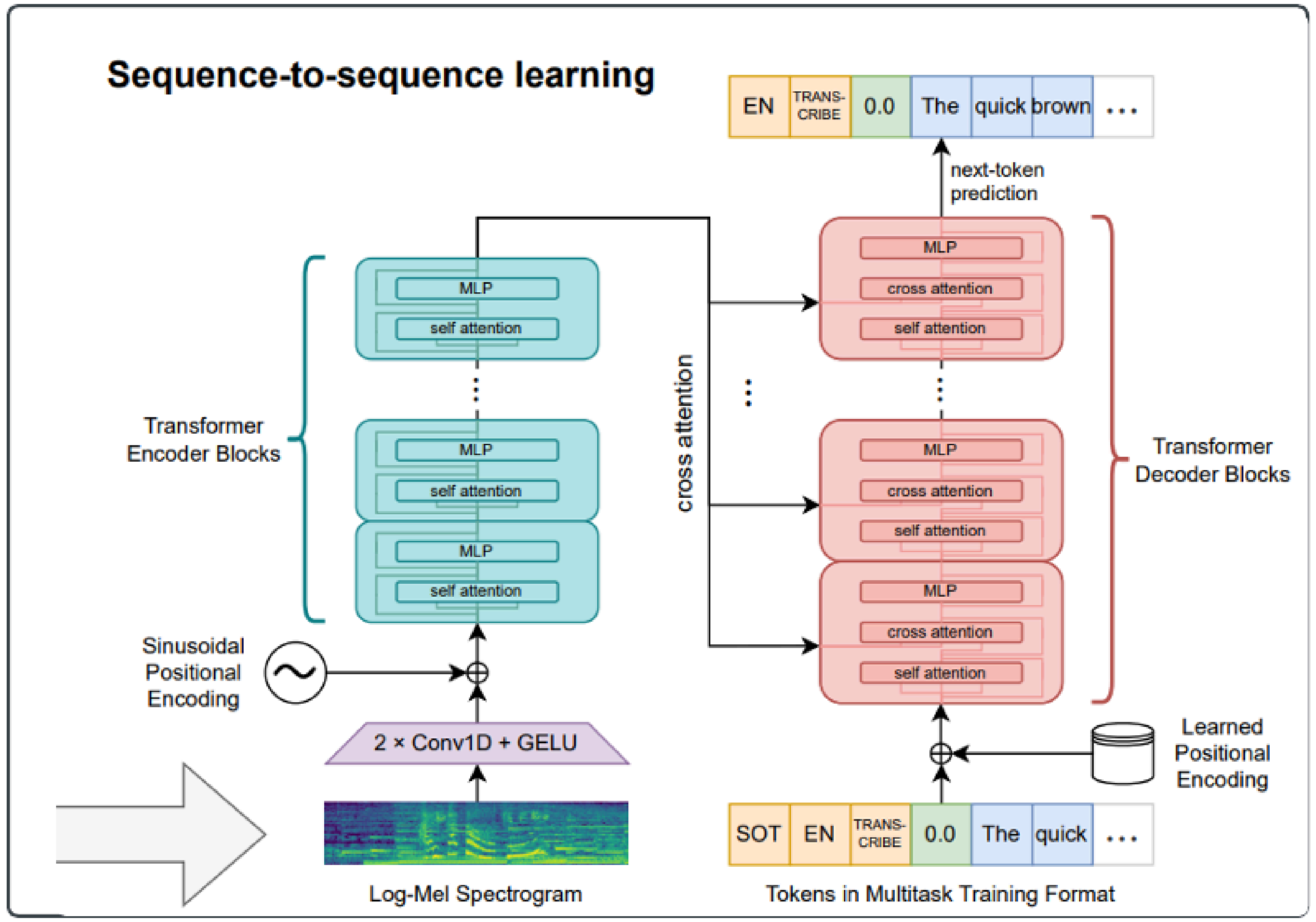
Pre-process audio and transcription followed by padding

Set training arguments

Train Model using Trainer API

Evaluate model using WER (Word Error Rate), CER Rate(Character Error Rate) and Levenshtein Edit Distance

Open AI's whisper architecture


Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
...
next-token prediction
MLP
cross attention
self attention
Transformer Encoder Blocks
cross attention
Transformer Decoder Blocks
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Learned Positional Encoding
SOT
EN
TRANS-CRIBE
0.0
The
quick
...
Log-Mel Spectrogram
Tokens in Multitask Training Format

Source: OpenAI Whisper Architecture Diagram (https://openai.com/research/whisper)

Key challenges and Takeaways:

- *Tokenization issues caused by sequence-to-sequence model contributes to WER as well, confusing noise and silence for speech .*
- *MPS backend was caused by large batch sizes taken during hyperparameter tuning and Apple silicon GPUs have limited VRAM.*
- *The training process lasted atleast 5-6 hours.*
- *Wav2Vec model has large parameters and required more finetuning.*
- *Longform audio (30s-60s) dealt with tokenization issues.*

Dataset Details

Sources

- Collected speech samples from native speakers.
- Open source speech data like Mozilla, OpenSLR and Indic TTS

Features

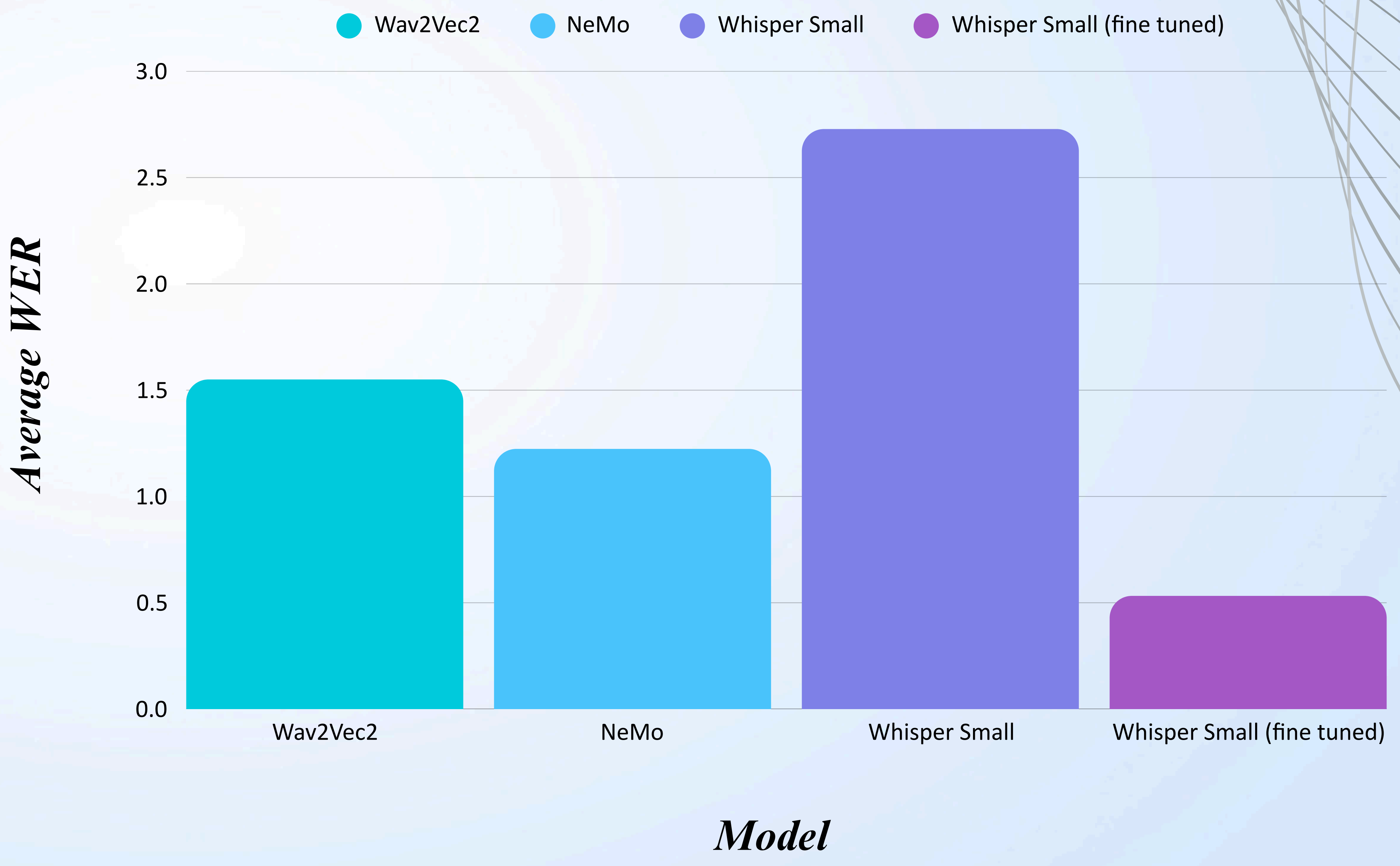
- Crowdsourced, clean read speech
- Noisy, conversational speech
- Studio-recorded, high-fidelity speech

Evaluation Metrics

- Word Error Rate (WER)
- Character Error Rate (CER)
- Levenshtein Distance

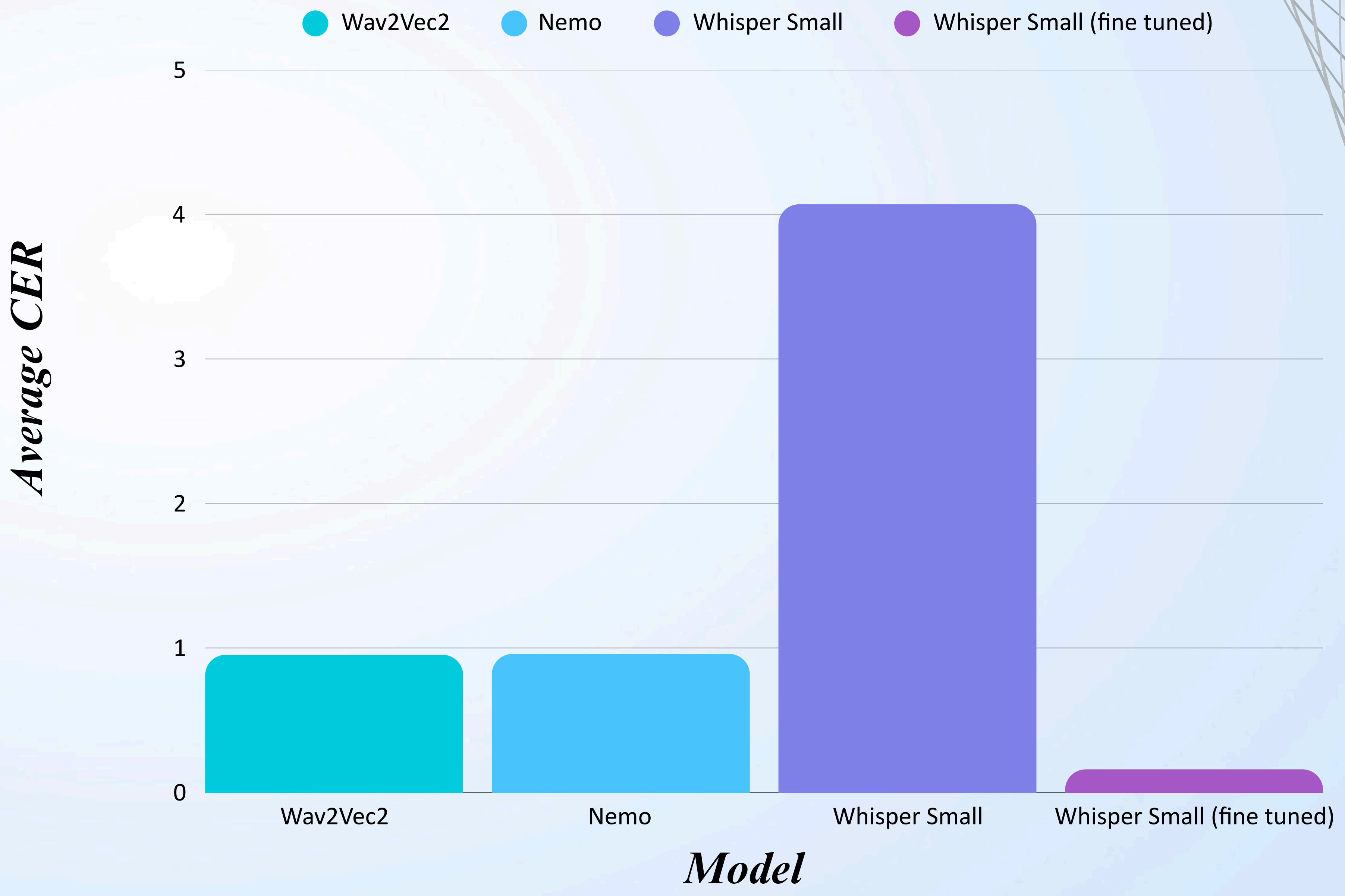
Results and Analysis

Model	Average WER
Wav2Vec2	1.5503
NeMo	1.2237
Whisper	2.7282
Whisper(Fine Tuned)	**0.5324**



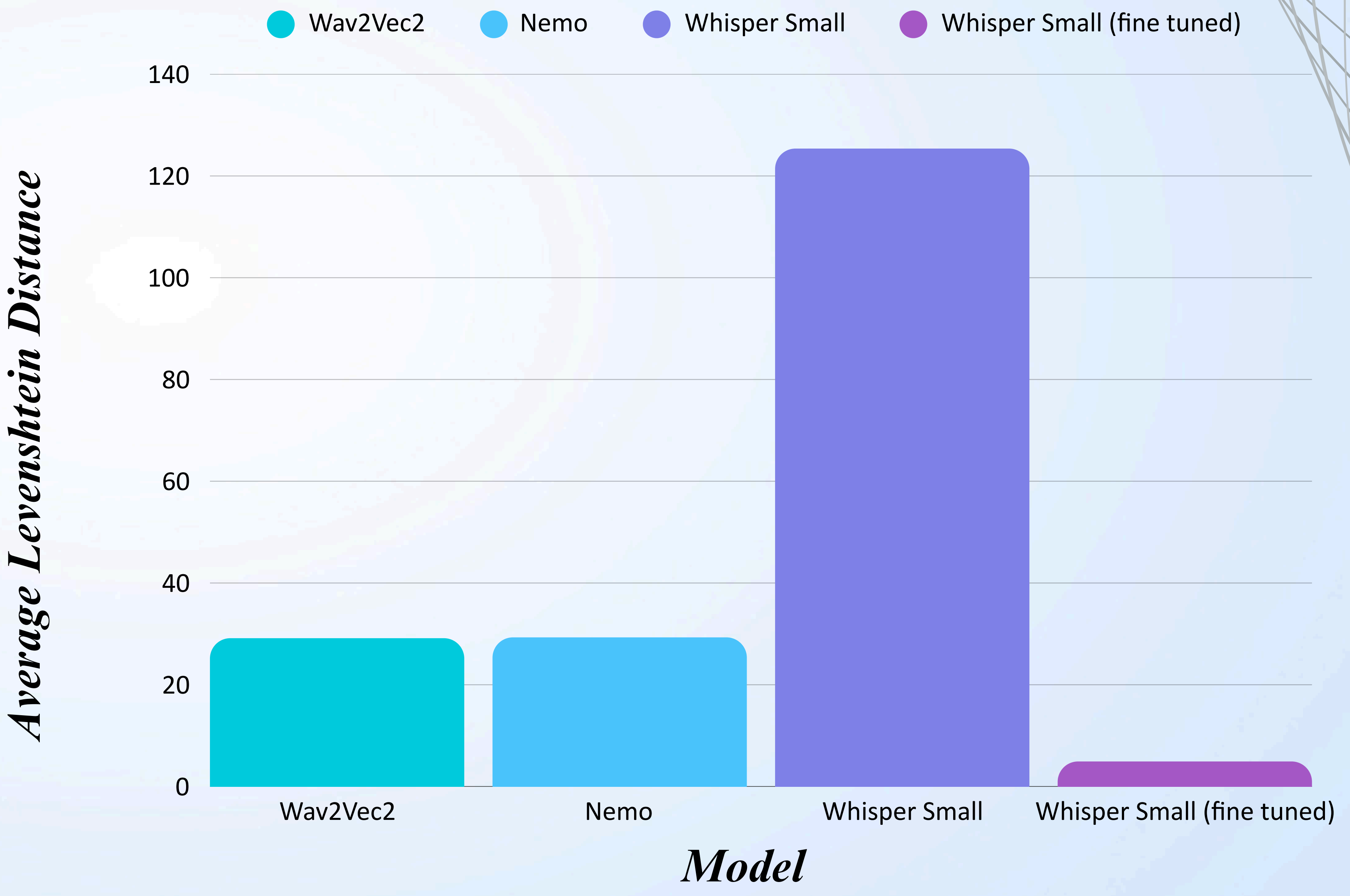
Results and Analysis

Model	Average CER
Wav2Vec2	0.9529
NeMo	0.9586
Whisper	4.0702
Whisper(Fine Tuned)	0.1599



Results and Analysis

Model	Average Levenshtein Distance
Wav2Vec2	29.1518
NeMo	29.3125
Whisper Small	125.3795
Whisper Small(Fine Tuned)	**4.9241**



- Integrate advanced Language Models (LLMs) to enhance contextual understanding, correct grammar, and generate more fluent and natural Telugu transcriptions.
- Expand the model's training to cover all phonetic sounds present in the Telugu language, ensuring improved recognition accuracy and linguistic completeness.
- Develop a user-friendly mobile application to enable real-time Telugu speech-to-text conversion, making ASR accessible to a wider audience on the go.
- Optimize and deploy the ASR model on edge devices to enable low-latency, offline Telugu speech recognition in resource-constrained environments.

THANK YOU

FOR YOUR ATTENTION