Enhancing Urdu ASR with Whisper v3: Fine-Tuning on Latest Datasets and Realistic Multi-Speaker Evaluation with SLM Post-Processing

Zehra Ahmed*, Farah Inayat*, Zuha Aqib*, Sajjad Haider
Department of Computer Science
Institute of Business Administration
Karachi, Pakistan
{z.ahmed.26965, f.inayat.26912, z.aqib.26106}@khi.iba.edu.pk, sahaider@iba.edu.pk

Abstract

Automatic Speech Recognition (ASR) for low-resource languages like Urdu remains challenging due to limited training data, dialectal variations, and orthographic inconsistencies. While prior efforts have fine-tuned earlier Whisper models (e.g., v2), this work evaluates the latest OpenAI Whisper v3-large and v3-large-Turbo variants, fine-tuned on the most recent CommonVoice (23), Fleurs and CSaLT datasets, comprising over 31 hours of Urdu speech. To enhance realism, we assess performance not only on standard benchmarks but also on a novel evaluation set derived from YouTube videos featuring dynamic two-person discussions between news anchors and on-site reporters, capturing natural interruptions, accents, and background noise. Additionally, we integrate Urdu-capable Small Language Models (SLMs) as a post-processing layer to correct spelling errors and normalize transcriptions, addressing common ASR pitfalls in Urdu script. SLM experiments demonstrate 11% WER reduction compared to baseline vanilla Whisper v3-Turbo, while overall the fine-tuned v3-turbo model outperforms its counterpart in noisy, multi-speaker scenarios by 5.81%. Our contributions include updated benchmarks for modern Whisper models on contemporary datasets, a realistic multi-speaker evaluation protocol, and an SLM-enhanced pipeline that improves transcription accuracy by 5.05%. This advances Urdu ASR toward practical deployment in media and communication applications.

Keywords: Urdu Automatic Speech Recognition (Urdu ASR), LoRA fine-tuning (PEFT), Real-world multi-speaker evaluation

1. Introduction

Over the past ten years, there have been significant advancements in Automatic Speech Recognition (ASR), mainly due to developments in deep learning, transformer-based sequence modeling, and large-scale multilingual training. For many high-resource languages, modern ASR systems can now produce transcriptions with accuracy close to that of a human. These systems are used extensively in voice assistants, real-time transcription, and human computer interaction systems.

However, not all languages experience the same improvements in ASR performance as low resource languages continue to be severely underserved. This imbalance has resulted in a clear gap between available resources and system performance in global ASR development. Languages with limited resources often suffer from the absence of standardized datasets, mature linguistic tools, and sustained research focus which restricts both effective model training and meaningful evaluation. Urdu is a notable example of this inequality, which remains underrepresented in both ASR research and commercial applications. Factors such as scarce annotated data, complex sound and writing structures, and a lack of proper language tools make the situation even more difficult. Because of this, the current multilingual ASR systems do not work well for the Urdu language.

In addition, despite being the national language of Pakistan, for a large portion of Pakistani speakers, Urdu is their second language, leading to significant influence from regional mother tongues (such as Punjabi, Pashtun, and Sindhi) that manifest as heavy accents and dialects, which make speech modeling harder. In addition, in everyday conversations, especially in urban educated contexts, speakers often mix Urdu with English but this code mixing does not follow fixed spelling rules creating confusion during transcription. Spoken Urdu also commonly drops or softens sounds at the end of words or sentences which results in differences between how words are written and how they are spoken. These issues, when combined, increase the gap between training data, which is usually clean and carefully read, while speech and real-world audio contain noise, interruptions, and overlapping speakers.

Earlier evaluation studies for Urdu ASR report high Word Error Rates (WERs) and frequent pronunciation and word-level mistakes. This shows that relying only on zero-shot multilingual models is not enough to achieve reliable Urdu speech recognition. While recent studies have laid important groundwork for Urdu ASR and have provided valuable baselines and demonstrated the feasibility of Urdu ASR, several opportunities for extension remain. Prior work has primarily focused on evaluating lightweight or zero-shot models, often using read speech recorded in quiet settings. Arif et al. (2025) presented the first comprehensive

* These authors contributed equally to this work.

benchmark of Urdu ASR models, evaluating Whisper, MMS, and Seamless-M4T on both read and self-introduced conversational speech datasets. Their results show that Seamless-Large performs best on read speech, while Whisper-Large achieves superior performance on conversational data. The study highlighted persistent challenges in low-resource Urdu ASR, particularly orthographic variation, code-switching, and the limitations of WER as a sole evaluation metric, emphasizing the need for robust text normalization.

Another work by Sehar et al. (2025) evaluated Whisper's zero-shot and few-shot performance on Pashto, Punjabi, and Urdu, three low-resource South Asian languages. While Whisper-Large performed comparatively better in zero-shot settings, it still struggled with script inconsistencies, dialectal variation, and phonetic complexity. The study demonstrated that few-shot fine-tuning (up to 15 hours of data) significantly reduced WER, highlighting the effectiveness of domain-specific adaptation for improving ASR performance in low-resource contexts.

Despite these important contributions, certain aspects remain underexplored. The conversational speech evaluated in prior work, while valuable, typically consists of controlled monologues rather than the dynamic, multi-speaker interactions found in media, everyday communication, or field reporting. Similarly, while post-processing strategies have shown promise for improving transcript quality in other low-resource contexts, such as Znotins et al. (2025), their application to Urdu, particularly for addressing orthographic variation and spelling inconsistencies in long-form conversational speech, remains relatively unexplored.

Furthermore, recent advances have created opportunities to extend existing benchmarks. OpenAI's release of Whisper Large-v3 and Large-v3-Turbo, along with substantial updates to multilingual datasets like Mozilla Common Voice v23, have not yet been systematically evaluated for Urdu. This presents an opportunity to build upon the foundations laid by Arif et al. and Sehar et al. by establishing updated baselines, introducing a more realistic multi-speaker evaluation protocol, and exploring SLM-based correction as a potential next step for the community.

Additionally, the datasets commonly used are vastly different from real-world audio in which such models need to be deployed. They usually have an average duration of 15 seconds, are monologues, and include unnatural pauses between words. Hence, this lack of benchmarks makes it difficult to assess progress, compare systems, or understand how modern multilingual ASR models perform on real-world Urdu speech.

To address this gap, the paper focuses on establishing strong reproducible and comparative baselines for Urdu ASR using OpenAI's Whisper Large-v3 and Whisper Large-V3-Turbo and the impact of fine-tuning these models on the latest Urdu speech datasets, such as CommonVoice23. Additionally, we evaluate our experimental models on real-world two-person noisy audio datasets curated from news segments on YouTube. And finally, we measure the effectiveness of decoder level optimizations and post transcription cleanup using Small Language Models to reduce real-world transcription errors.

The rest of the paper is organized as follows. Section 2 walks through the experimental design, while Section 3 discusses our key findings and results. Finally, Section 4 concludes the paper and provides future research directions.

All code and model outputs are publicly available on GitHub.[1]

2. Design of Experiments

2.1 Research Questions

The experimental setup is designed to answer the following research questions:

- **RQ1:** How does Whisper large-turbo compare to other Whisper models on Urdu ASR in zero-shot, fine-tuned, and real-world settings?
- **RQ2:** Does fine-tuning on curated audio datasets improve or degrade model performance on real-world audio inferences?
- **RQ3:** Can decoder-level optimization reduce WER at inference time without retraining?
- **RQ4:** Do Language Model-based post-processing methods further reduce error rate on long-form, contextual Urdu speech?

2.2 Replicability and Experimental Setup

To ensure replicability and a fair comparison across experiments, all models, datasets, and evaluation procedures are trained and evaluated

[1] https://github.com/awaaz-se-alfaaz-fyp/Enhancing-Urdu-ASR-with-Whisper-v3

under controlled hardware and software environments.

2.2.1 Hardware Specifications

All experiments were conducted on a GPU-accelerated High-Performance Computing Cluster accessed via Ubuntu Virtual Machine. The Ubuntu VM ran Ubuntu 22.04.5 LTS (Jammy), with 32 vCPUs (Intel Xeon) available, 62 GiB system memory, and 1x NVIDIA A40 48GB GPU.

2.2.2 Software Stack

Experiments were implemented in PyTorch using the Hugging Face Transformers ecosystem with GPU acceleration and mixed precision training. The training pipeline included Whisper processor and sequence-to-sequence generation model, parameter-efficient fine-tuning via LoRA, mixed precision (FP16) optimization and AdamW optimizer. Random seeds were fixed across Python, NumPy, and PyTorch to ensure deterministic behavior. All dataset shuffling, subsampling, and train-test construction followed the same seed configuration.

2.2.3 Fixed Controls

To guarantee a fair comparison, several factors were kept the same across all experiments. Train/validation split logic, text normalization rules, audio preprocessing steps, evaluation datasets, and metric calculation scripts were identical for all models regardless of their architecture or hyperparameters. Decoding settings were also fixed within each experiment.

2.3 Models

The paper focuses mainly on two OpenAI models: Whisper large-v3 and Whisper large-turbo. Their selection is grounded in prior Urdu ASR benchmarking work, particularly Arif et al. (2025), which shows Whisper-large variants achieving the strongest performance on conversational Urdu speech. While large-v3 represents the highest-capacity model in the Whisper family, large-turbo offers a practical efficiency trade-off, being approximately 4× smaller and 4× faster while maintaining comparable accuracy. At the time of this study, no published benchmarks or fine-tuning results existed for Whisper large-turbo on Urdu, making its systematic evaluation a core contribution of this work.

Both models follow an encoder-decoder Transformer architecture. The encoder processes log-Mel spectrogram representations or audio inputs, while the decoder generates text tokens autoregressively.

Both models are evaluated under three configurations:

- **Zero-shot:** uses pretrained weights. The evaluation pipeline, however, specifies the task as transcription and the language as Urdu while evaluating.
- **Fine-tuned:** using LoRA on Urdu Speech datasets described in section 2.5.1
- **Real-world:** The models with the best parameter configurations for both large-v3 and turbo were selected and evaluated on two-person noisy YouTube evaluations.

Additionally, for our post-processing SLM evaluations, we evaluated 4 models with either multilingual capabilities or models specifically fine-tuned for Urdu tasks and datasets. These include:

(a) Qwen3-14B (Release Date: April 28, 2025)
(b) tiny-aya-fire (Release Date: February 17, 2026)
(c) Qalb-1.0-8B-Instruct (Release Date: January 13, 2026)
(d) Gemma-2-9b (Release Date: June 27, 2024)

All models were taken from HuggingFace using Transformers.

2.4 Data and Pre-processing

2.4.1 Datasets

We use three publicly available Urdu speech datasets for training and evaluating for zero-shot and fine-tuned settings:

(a) Mozilla Common Voice (Urdu, V23)
(b) FLEURS (Urdu subset)
(c) CSaLT Urdu dataset (Arif et al., 25)

Together, these datasets provide diverse speakers and recording conditions but share several structural characteristics. Most utterances are short-duration clips (typically ≤15 seconds), recorded in relatively clean environments, and feature single-speaker read speech with noticeable pauses between words. While suitable for benchmarking, these characteristics differ significantly from real-world conversational Urdu speech.

To evaluate real-world performance, we created a novel evaluation set from YouTube videos featuring dynamic two-person discussions between news anchors and on-site reporters. These clips include natural interruptions, accents, background noise, and are significantly longer (average 80 seconds) than training or benchmark audio. This Real-World

Multi-Speaker dataset reflects realistic broadcast conditions beyond short, clean, single-speaker speech.

While selecting videos, we avoided political, religious, or figure-specific content and ensured minimal speaker overlap. Due to copyright restrictions, the dataset was used only for evaluation and not for fine-tuning.

2.4.2 Data Mixing and Train/Test/Val Split

All three datasets were first standardized into a unified metadata format containing audio paths, normalized transcripts, speaker identifiers (if available), and duration information. To avoid dataset bias, samples from each corpus were randomly shuffled and proportionally mixed before splitting.

The combined corpus was divided into:

- Training set: 80% of total samples
- Validation set: 10% of total samples
- Test set: 10% of total samples

A fixed random seed was used to ensure reproducibility across experiments.

2.4.3 Text Normalization

The Urdu transcripts collected from different datasets were not fully consistent. They differed in punctuation, number formats, spacing, and Unicode characters. To fix this, a single normalization process was used for all data. This process included:

- Converting all text into standard Unicode Urdu script
- Removing unnecessary punctuation and special symbols
- Making Arabic and Urdu character forms consistent
- Changing numbers into their spoken Urdu word form
- Removing extra spaces at the beginning and end of the text
- Applying lower-case formatting where needed for tokenizer consistency

No heavy language processing techniques, such as stemming or word reduction, were used. This was done to keep the original meaning of the text intact for accurate ASR evaluation.

2.4.4 Audio Pre-processing

Before training, all audio files were converted into the same format to keep the data consistent. Each audio file was changed to a mono channel, a 16 kHz sampling rate, and a 16-bit PCM waveform. Extra silence at the beginning and end of the recordings was removed, while normal pauses within speech were kept. Audio clips that were longer than the model's allowed length were split into slightly overlapping segments so that no speech was lost during training.

For feature extraction, log-Mel spectrograms were used following Whisper's standard preprocessing steps. This ensured the audio features were fully compatible with the model's pretrained encoder.

2.5 Fine-Tuning setup

Fine-tuning was performed using the AdamW optimizer with a learning rate of 1e-4 to 1e-6, batch size 8, and 8–12 epochs. Mixed-precision (FP16) training was enabled for computational efficiency, label length was capped at 128–256 tokens for stability, and a learning rate scheduler was used to mitigate overfitting.

For LoRA, Turbo used rank 32, alpha 64, and dropout 0.05 applied to the query, key, value, and output projection layers, while Large used rank 16 and alpha 32 with other settings unchanged. Hyperparameters were chosen based on stability checks and hardware constraints.

2.6 Benchmark Evaluation

For benchmark evaluation, we used the pooled dataset described in Section 2.5, constructed by proportionally mixing Mozilla Common Voice v23 (Urdu), FLEURS (Urdu), and CSaLT before applying the 80–10–10 train/validation/test split. Fixed random seeds were used across all experiments to ensure consistency. Evaluation was performed on the held-out test portion of this pooled dataset using identical preprocessing, normalization, and scoring procedures described earlier. Word Error Rate (WER) was computed for each model under two configurations: (a) zero-shot and (b) fine-tuned (LoRA). We also applied a decoder sweep for Turbo.

2.7 Decoder Level Optimization

In addition to model training, we perform decoder-level optimization at inference time. A decoder sweep was conducted by varying decoding parameters such as beam width and length penalty to identify configurations that minimize WER on validation data.

Decoder optimization was applied uniformly across the baseline and fine-tuned Turbo model and did not involve retraining. This allowed us to quantify inference-time improvements achievable through decoding strategy alone.

2.8 SLM-Based Post-Processing

While short benchmark utterances provide limited linguistic context, long-form news recordings enable contextual correction. For this reason, we applied a post-transcription cleanup step using Urdu-capable Small Language Models (SLMs) ranging from 8B to 14B parameters. We restricted corrections strictly to minimal character-level replacements without stylistic rewriting to preserve spoken Urdu characteristics.

These SLMs operated solely on outputs from the fine-tuned Turbo model and were used only during evaluation. A single consistent prompt was applied across all models which is given below.

You are an Urdu ASR error correction expert. Your ONLY task is to replace incorrectly transcribed Urdu words with their correct forms.
CRITICAL RULES:
- ABSOLUTELY NO punctuation (no ؟ ، - ! . , : ; " ' or any symbols)
- NO new words or phrases
- NO reordering
- NO grammar changes
- ONLY replace wrong words with correct ones
- If unsure about a word, leave it unchanged
Think of this as a word-by-word dictionary replacement, not a rewrite.
Fix ONLY the incorrectly transcribed words in this Urdu text. Replace wrong words with correct ones based on context. Add NO punctuation.
Urdu text: {text}
Corrected text:

2.9 Metrics

Performance is measured using metrics from the jiwer library. All transcriptions were evaluated using identical normalization and scoring procedures. Word Error Rate (WER) was the only metric used for fine-tuning and Youtube evaluations, and constructing benchmarks for the latest datasets. For SLM evaluation, WER, along with Character Error Rate (CER), Match Error Rate (MER), and Word Information Lost (WIL) were used.

3. Results & Findings

This section reports benchmark Word Error Rate (WER) results for Whisper large-v3 and Whisper large-v3-turbo under three settings: (i) base (zero-shot), (ii) LoRA fine-tuned, and (iii) decoder sweep (only for Turbo). All results use the same normalization and scoring pipeline described in Section 2.10.

3.1 Benchmark Evaluation

Table 1 summarizes WER on the three benchmark corpora used in this study. For each dataset, the “Base” column corresponds to the pretrained model evaluated with Urdu transcription settings, “Fine-tuned” corresponds to the best LoRA run trained on the Urdu training data described in Section 2.5, and “Decoder sweep” row corresponds to the best inference-time decoding configuration identified through the sweep described in Section 2.8. WER values are reported as percentages (lower is better).

	CSaLT		Common Voice		FLEURS	
Model	Base	Fine-tuned / Decoder Sweep	Base	Fine-tuned / Decoder Sweep	Base	Fine-tuned / Decoder Sweep
whisper-large -v3	23.78	23.93	30.4	30.18	22.35	23.33
whisper-large -v3-turbo **FINE-TUNED**	31.11	25.52	39.66	31.32	22.83	22.81
whisper-large -v3-turbo **DECODER SWEEP**	-	25.39	-	32.27	-	21.63

Table 1. Benchmark WER (%) for Whisper large-v3 and large-v3-turbo using finetuning

Several consistent trends emerge:

- Turbo benefits strongly from LoRA fine-tuning on CSaLT and Common Voice. Whisper large-v3-turbo improves on CSaLT (31.11 → 25.52) and on Common Voice (39.66 → 31.32), indicating that parameter-efficient adaptation substantially reduces its weaker zero-shot performance on these benchmarks.
- Large-v3 is comparatively stable on CSaLT and Common Voice under this fine-tuning setup. On CSaLT, large-v3 remains essentially unchanged (23.78 → 23.93). On Common Voice, it improves marginally (30.40 → 30.18), suggesting limited gains from the same fine-tuning recipe.
- On FLEURS, gains are not consistent across models. Large-v3 slightly worsens (22.35 → 23.33), while turbo shows a very small improvement (22.83 → 22.81). This indicates that the fine-tuning configuration used here does not uniformly transfer to FLEURS.

3.2 Decoder-Level Optimization

Decoder sweeps quantify improvements achievable without retraining by varying decoding parameters and selecting the best configuration.

- On CSaLT, turbo improves slightly from 25.52 (fine-tuned) to 25.39 (decoder sweep), showing a small but consistent decoding gain.
- On Common Voice, turbo improves from 31.32 (fine-tuned) to 32.27 (decoder sweep) — i.e., decoding worsens in this sweep result, so it should be reported as no improvement (or as a degradation), not as a benefit.
- On FLEURS, turbo improves from 22.81 (fine-tuned) to 21.63 (decoder sweep), indicating that decoding choices can provide a measurable gain on this benchmark.

Decoder-sweep results are not reported for large-v3 in Table 1, where the sweep was not available in the final experiment logs used for summarization.

3.3 Evaluating Real-world Multi-Speaker Audios

We selected 30 YouTube videos from several news channels, spanning a total duration of 40 mins. These were news segments containing two speakers taking turns, background noise, and interruptions, with each clip having an average length of 80 seconds. We evaluate both zero-shot and fine-tuned Whisper Large-V3 and Whisper Large-Turbo to gauge effectiveness of these models in a real-world deployed setting. The results are summarized in Table 4.

Although both models perform competitively on the benchmark datasets, real-world conversational audio exhibits greater variability. While the overall average WER on the YouTube dataset remains comparable to benchmark results, per-video analysis reveals noticeable fluctuations, particularly in segments featuring stronger regional accents or rural reporting contexts. Fine-tuning large-v3 does not produce substantial changes in real-world performance, whereas fine-tuning large-turbo yields a moderate improvement of 5.81%. In zero-shot settings, turbo demonstrates slightly lower robustness than large-v3 under noisier conversational conditions.

Model	WER
v3-large	24.27%
v3-large fine-tuned	24.48%
v3-turbo	28.58%
v3-turbo fine-tuned	26.92%

Table 4: Real-World Multi-Speaker Performance on YouTube Dataset

3.4 SLM cleanup as a Post Processing step

Given that each audio file in this YouTube dataset is considerably longer than the datasets described in section 2.5.1, they are able to hold context. We used this characteristic to introduce SLM cleanup as a post-processing step to make minor fixes to spellings. We use our best-performing fine-tuned Turbo model to generate transcribed outputs and pass them to the four selected SLMs mentioned earlier.

A key challenge in using SLMs as a post-processing step was maintaining an appropriate balance between correction and preservation. On one end, excessive correction risked transforming conversational, spoken Urdu into overly formal, grammatically refined text, replacing slang, regional expressions, or code-switched words with standardized equivalents and thereby distorting the original speech style and spiking WER. On the other end, overly conservative decoding resulted in

minimal or no corrections, effectively reproducing the input transcript without meaningful improvement. Careful prompt design and parameter tuning were therefore required to preserve spoken nuances while allowing targeted spelling and recognition corrections.

For additional comparison, we also evaluated GPT-OSS-20B as a larger LLM baseline. However, we observed that instruction-tuned SLMs with chat templates and direct text-generation objectives performed more reliably for transcript cleanup than reasoning-oriented models such as GPT-OSS. We also conducted exploratory comparisons using larger proprietary LLMs (e.g., Gemini and Grok variants, ~70B+ scale) through publicly available interfaces. These models demonstrated substantially stronger post-processing performance on the same transcripts, in some cases yielding near single-digit WER reductions. This performance gap suggests that closed, production-grade systems may benefit from additional optimization layers, proprietary fine-tuning, or system-level integration that are not directly comparable to open-weight models evaluated in this study. However, due to limited transparency and reproducibility constraints, these systems were not included in the formal benchmark analysis. Importantly, this comparison indicates that open-weight SLMs still have considerable headroom for improvement in Urdu transcript correction tasks, and that targeted fine-tuning or alignment specifically for conversational Urdu may yield further gains, presenting a promising direction for future work. The findings also suggest that conversational fine-tuning and generation alignment are more critical for post-ASR cleanup than raw parameter scale.

After post-processing, we evaluated transcript quality using four metrics from the jiwer library: Word Error Rate (WER), Character Error Rate (CER), Match Error Rate (MER), and Word Information Lost (WIL). Among all tested models, Qwen3-14B consistently achieved the strongest improvements, outperforming even Qalb-1.0-8B-Instruct and Tiny-Aya-Fire, which are specifically designed for Urdu and South Asian languages.

Error analysis revealed that regional accents and dialectal pronunciations contributed significantly to higher WER, particularly in cases where colloquial or non-standard lexical items, common in everyday speech, were transcribed into non-dictionary forms. A detailed per-file inspection of the YouTube dataset further showed variability in SLM effectiveness: several long-form audio files exhibited substantial WER reductions after cleanup, while others experienced marginal degradation compared to the original ASR output. This highlights that SLM-based post-processing is context-dependent and influenced by conversational continuity, speaker variability, and dialectal richness. In particular, SLM correction was particularly effective when contextual continuity existed across sentences, and regional accents were low. Additionally, some WER inflation was attributable to inconsistencies between model outputs and gold transcripts rather than purely recognition errors. As also highlighted by Arif et al. (2025), Urdu orthographic variation, particularly the joining or separation of words in written form, introduces systematic evaluation discrepancies. We observed multiple cases where semantically correct transcriptions were penalized due to differences in spacing conventions or compound word segmentation. This reinforces the need for standardized normalization practices when benchmarking Urdu ASR systems.

We compare the improvements across 4 metrics for the 4 models and present our findings in Table 5, where using Qwen/Qwen3-14B as a post-processing step reduced WER by 5.05% on long-form conversational data. A qualitative example is shown in Table 6.

Metrics	WER	CER	MER	WIL
v3-turbo	27.09%	12.87%	25.72%	39.05%
v3-turbo fine-tuned	25.27%	11.79%	23.94%	36.76%
Tiny Aya Fire	26.67%	13.27%	25.42%	38.50%
Qalb8B	68.35%	58.11%	67.03%	75.94%
Gemma9b	54.56%	46.41%	53.99%	61.96%

Qwen3-14B	**24.00%**	**11.77%**	**22.71%**	**34.72%**

Table 5: SLM Post-Processing Results (YouTube Dataset)

Pre SLM	Raw ASR	Post SLM	New ASR
20.71%	سابق کپتان رمیز راجہ کراچی کنگز کے نوجوان کھلاڑیوں سے متاثر کہتے ہیں کہ کراچی کنگز کے ٹیم میں باسلاحیت نوجوان کھلاڑی شامل ہیں قاسم اکرم، ارشد اقبال، ابباس آفریڈی اور محمد حارس لمبے عرصے تک کراچی کنگز کے لیے کھیل سکتے ہیں کنگز کے پاس ینگ پلیز ہیں جو آپ کی توجہ کا مرکز بن سکتے ہیں آگے چل قاسم اکرم، گوڈ لوکنگ بیٹسمن ہے لیک سائٹ پر جو ان کی بیٹنگ ہے وہ ویرات کولی کی آپ کو یاد آ جاتی ہے جیسے یہ فلک کرتے ہیں بال کو کنگز کے پاس دو ینگ اچھے فاس بولرز بھی ہیں عباس آفریدی اور عرشد اقبال لمبے کٹ کے ہیں سپیڈ بڑی اچھی یا ہارڈ لینڈ کو ہٹ کرتے ہیں کسی بھی موقع پہ آکے یہ آپ کو بولنگ کر سکتے ہیں اور عباس آفریدی نئے بال پہ یورکر کروا سکتے ہیں بہت اچھے دو یہ ینگ فاس بولرز ہیں جو کہ آگے کراچی کنگز کو کئی سال تک یہ سروس پروائیڈ کر سکتے ہیں کراچی کنگز کی یہ ایک اور اچھی ینگ پک ہے ان کا نام ہے محمد حارس بہت اچھی پوٹینچل ہے یہ بھی ٹیمپرومنٹلی سٹرونگ ہے اور کالٹیز بے انتہاں ہیں موسیقی	16.67%	سابق کپتان رمیز راجہ کراچی کنگز کے نوجوان کھلاڑیوں سے متاثر کہتے ہیں کہ کراچی کنگز کے ٹیم میں باصلاحیت نوجوان کھلاڑی شامل ہیں قاسم اکرم ارشد اقبال ابباس آفریدی اور محمد حارس لمبے عرصے تک کراچی کنگز کے لیے کھیل سکتے ہیں کنگز کے پاس ینگ پلیئرز ہیں جو آپ کی توجہ کا مرکز بن سکتے ہیں آگے چل قاسم اکرم گوڈ لوکنگ بیٹسمن ہے لیک سائٹ پر جو ان کی بیٹنگ ہے وہ ویرات کولی کی آپ کو یاد آ جاتی ہے جیسے یہ فلک کرتے ہیں بال کو کنگز کے پاس دو ینگ اچھے فاس بولرز بھی ہیں عباس آفریدی اور ارشد اقبال لمبے کٹ کے ہیں سپیڈ بڑی اچھی یا ہارڈ لینڈ کو ہٹ کرتے ہیں کسی بھی موقع پہ آکے یہ آپ کو بولنگ کر سکتے ہیں اور عباس آفریدی نئے بال پہ یورکر کروا سکتے ہیں بہت اچھے دو یہ ینگ فاس بولرز ہیں جو کہ آگے کراچی کنگز کو کئی سال تک یہ سروس پروائیڈ کر سکتے ہیں کراچی کنگز کی یہ ایک اور اچھی ینگ پک ہے ان کا نام ہے محمد حارس بہت اچھی پوٹینچل ہے یہ بھی ٹیمپرومنٹلی سٹرونگ ہے اور کوالٹیز بے انتہاں ہیں موسیقی

Table 6: Qualitative Example

4. Conclusion

This work established updated, reproducible baselines for Urdu ASR using Whisper Large-v3 and Large-v3-Turbo, addressing a critical gap in benchmarking for this underserved language. Our results demonstrate that LoRA fine-tuning consistently benefits Whisper Large-v3-Turbo. Furthermore, integrating Small Language Models as a post-processing step shows measurable potential for reducing transcription errors, though this remains an underexplored direction. The development of Urdu-specific SLMs, tailored to the language's orthographic and phonetic variability, represents a promising avenue for future work.

5. Ethics

All training datasets used in this study (Mozilla Common Voice 23 (Urdu), FLEURS (Urdu subset), and the CSaLT Urdu dataset) are publicly available under research permissible licenses and were used in accordance with their respective terms. We did not collect or introduce any private or personally identifiable data. Real-world evaluation audio was curated from publicly accessible YouTube news segments and used strictly for research evaluation purposes only. These recordings were not used for training, redistributed, or released, in order to respect copyright and platform policies. All evaluated ASR models are publicly released, and our preprocessing, fine-tuning (LoRA), and decoding procedures are documented to support transparency and reproducibility.

Additionally, while SLM-based post-processing was restricted to minimal orthographic corrections, language models may implicitly favor standardized forms of Urdu over colloquial variants. We emphasize that this work is intended to improve accessibility and low-resource language research, and we advocate for responsible deployment that respects privacy, fairness, and human rights considerations.

6. References

Arif, S., Khan, A. J., Abbas, M., Raza, A. A., and Athar, A. (2025). WER we stand: Benchmarking Urdu ASR models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5952–5961, Abu Dhabi, UAE. Association for Computational Linguistics.

Sehar, N. U., Khalid, A., Adeeba, F., and Hussain, S. (2025).Benchmarking Whisper for low-resource speech recognition: An n-shot evaluation on Pashto, Punjabi, and Urdu. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 202–207, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Znotins, A., Gruzitis, N., and Dargis, R. (2025).*From conversational speech to readable text: Post-processing noisy transcripts in a low-resource setting.* In Proceedings of the Tenth Workshop on Noisy and User-generated Text, pages 143–148, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Qwen Team. (2025). Qwen3 Technical Report. arXiv preprint arXiv:2505.09388. https://arxiv.org/abs/2505.09388

Hassan, M. T., Ahmed, J., & Awais, M. (2026). Qalb: Largest state-of-the-art Urdu large language model for 230M speakers with systematic continued pre-training. arXiv preprint arXiv:2601.08141.https://doi.org/10.48550/arXiv.2601.08141

Gemma Team. (2024). Gemma. Kaggle. https://doi.org/10.34740/KAGGLE/M/3301

Cohere Labs. (2024). Tiny Aya Fire. Hugging Face. https://huggingface.co/CohereLabs/tiny-aya-fire

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). *Robust speech recognition via large-scale weak supervision*. Proceedings of the 40th International Conference on Machine Learning, 202, 28492–28518. https://proceedings.mlr.press/v202/radford23a.html

Arif, S., et al. (2025). CSaLT-Voice. Hugging Face. Deepfake Defense: *Constructing and Evaluating a Specialized Urdu Deepfake Audio Dataset.* https://huggingface.co/datasets/urdu-asr/csalt-voice

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., and Bapna, A. (2022). *FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech*. arXiv preprint arXiv:2205.12446.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). *Common Voice: A Massively-Multilingual Speech Corpus.* In Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), pages 4211–4215, Marseille, France. European Language Resources Association (ELRA).

Hanindhito, B., Patel, B., & John, L. K. (2025). Large language model fine-tuning with low-rank adaptation: A performance exploration. *Proceedings of the 16th ACM/SPEC International Conference on Performance Engineering (ICPE '25)*, 92–104. https://doi.org/10.1145/3676151.3719377

Prinos, K., Patwari, N., & Power, C. A. (2024). Speaking of accent: A content analysis of accent misconceptions in ASR research. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 1245–1254. https://doi.org/10.1145/3630106.3658969