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 (IBA), 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 variation, code-switching, and orthographic inconsistencies. This work evaluates OpenAI Whisper large-v3-Turbo for Urdu ASR and performs Low Rank Adaptation (LoRA) fine-tuning on Common Voice v23, FLEURS and CSaLT datasets, drawn from a pooled corpus of 31 hours of Urdu speech. To gauge robustness in the real world, we also created an evaluation set from YouTube that includes news segments with two people, namely an anchor and an on-site reporter, with natural accents, interruptions, and background noise. Apart from this, the paper explored Urdu-capable Small Language Models (SLMs) ranging from 3.5B to 14B parameters for generative error correction (GEC). Whisper-Turbo achieved 25.33% WER on the evaluation set from YouTube, whereas Whisper-Turbo with 4-bit quantized Gemma3-12B as the SLM GEC attained 21.75% WER, which translates to a 14.14% relative reduction. The results indicate that there is a potential application of SLM post-processing for long-form Urdu transcription, at the cost of extra runtime.

1 Introduction

Although ASR has significantly advanced in recent years, progress on low-resource languages remains limited (Ahlawat et al., 2025). One such language is Urdu, which is widely used in Pakistan and across South Asia, but the problem of reliable Urdu ASR is challenging because of the lack of annotated speech data, dialectal differences, code switching, and orthographic irregularities. Urdu is also affected by regional languages like Punjabi, Pashto and Sindhi native speakers, giving rise to accented and variable pronunciations among many people who speak Urdu (Afsheen et al., 2014). While this cross-linguistic influence is not explicitly annotated in our training or evaluation data, it motivates the need for ASR systems that are robust to accent variation in real-world Urdu speech. Additionally, everyday Urdu speech is also often interspersed with Urdu and English, and written Urdu does not have standardized spelling and spacing conventions, which makes it harder to evaluate and transcribe.

The recent studies have provided important baselines for Urdu and other South Asian languages. (Arif et al., 2025b) evaluated Urdu ASR models like Whisper, MMS, and Seamless-M4T, with the issues of orthographic variation, code-switching, and the inadequacy of WER as the sole evaluation metric. (Sehar et al., 2025) investigated Whisper on three languages: Pashto, Punjabi and Urdu, demonstrating that fine-tuning can enhance the performance of low-resource ASR models, but the dialectal and script-related issues still pose a challenge. Existing evaluations, however, are mostly conducted on short and cleaner texts, single speaker or controlled speech, and not on real-world Urdu audio with interruptions, background noise, accented speech, and rapid alternation of speakers.

This distinction is particularly relevant for media and communication settings, where Urdu speech often appears in long-form, conversational, and multi-speaker formats. While post-processing methods have improved transcript readability in a few low-resource contexts (Znotins et al., 2025), their use for Urdu ASR remains underexplored, particularly for correcting spelling and orthographic variation in long-form conversational speech. Recently, Large Language Models (LLMs) have shown strong performance as ASR post-processors through Generative Error Correction (hereafter GEC) [(Ma et al., 2023) , (Yamashita et al., 2025)]. This approach suits Urdu well because its Perso-Arabic script and complex morphology often cause homophone errors and orthographic ambiguities. However, prior work mostly relies on proprietary models such as

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 408–417
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_042

ChatGPT (Ma et al., 2023), which introduce API costs, latency, and privacy concerns that hinder offline or budget-constrained deployment. The emergence of capable open-weight Small Language Models (SLMs) presents an opportunity to perform GEC with lower computational overhead and without dependence on external APIs, offering a practical, locally-deployable alternative. This motivates our focus on evaluating open-weight SLMs rather than proprietary large-scale models for Urdu GEC.

To address these gaps, we evaluate OpenAI Whisper large-v3-Turbo (hereafter, Whisper-Turbo) (Radford et al., 2023) for Urdu ASR, applying Low Rank Adaptation (LoRA) fine-tuning across Common Voice v23 (Ardila et al., 2020), FLEURS (Conneau et al., 2023), and CSaLT (Arif et al., 2025a). To assess whether this performance generalizes to unscripted speech, we introduce a YouTube-based multi-speaker evaluation dataset featuring unconstrained conversational news broadcasts, directly addressing the lack of real-world benchmarks for Urdu. Finally, we explore SLM-based Generative Error Correction (GEC) to reduce residual transcription errors while evaluating the latency trade-offs of post-processing.

2 Design of Experiments

2.1 Research Questions

The conducted experiments address two primary research questions. First, we investigate whether Low-Rank Adaptation (LoRA) fine-tuning on curated Urdu datasets improves transcription performance on real-world, multi-speaker conversational audio. Second, we examine whether SLM-based GEC further reduces errors in long-form Urdu transcripts.

2.2 Experimental Setup

All experiments were conducted on an NVIDIA RTX 4080 GPU with 16 GiB VRAM. Models were implemented using PyTorch and Hugging Face Transformers with mixed-precision training. Fine-tuning used LoRA with AdamW optimization. To ensure fair comparison, we fixed random seeds across Python, NumPy, and PyTorch, and used identical train/validation/test split logic, audio preprocessing, text normalization, evaluation datasets, and metric calculation scripts across experiments.

2.3 Models

This paper focuses on OpenAI Whisper large-v3-Turbo (Whisper-Turbo) (Radford et al., 2023) for Urdu ASR. Prior Urdu ASR benchmarking has shown Whisper-large variants to be competitive on conversational Urdu (Arif et al., 2025b). Whisper-Turbo offers this same competitive performance with greater efficiency than the full Whisper large-v3 model, which is why we selected it. We initially evaluated Whisper large-v3 alongside Whisper-Turbo to confirm this choice; however, Whisper large-v3's zero-shot and fine-tuned WER trends closely matched those already reported for Whisper-large variants in prior work (Arif et al., 2025b), offering no additional insight beyond what Whisper-Turbo captured. We therefore report results for Whisper-Turbo only throughout this paper. Whisper-Turbo was evaluated in both zero-shot and LoRA fine-tuned settings. We publicly release our fine-tuned LoRA adapter on Hugging Face to support reproducibility.[1]

On the YouTube evaluation set, we additionally evaluated SLM-based GEC. In this context, we evaluated four open-weight SLMs with multilingual or Urdu-specific capabilities, using a mix of quantized and full-precision models to enable fair comparison on the same machine: (a) Gemma3-12B (4-bit quantized) (Gemma Team, 2025), (b) Qwen3-14B (4-bit quantized) (Qwen Team, 2025), (c) Tiny-Aya-Fire (Cohere Labs, 2024), and (d) Qwen3-4B (Qwen Team, 2025). Due to the hardware constraints of the machine described in Section 2.2, quantized versions of the larger models were used to enable local inference. We additionally evaluated Gemma3-12B at 8-bit quantization; however, as its performance differed only marginally from the 4-bit version, we retain and report results for Gemma3-12B (4-bit) only, to allow for a more consistent comparison across models. These SLMs were employed strictly as text-in-text-out correctors; they receive ASR transcripts as input without access to the original audio or gold references. To rigorously isolate the contribution of post-processing, we applied SLM correction to the outputs of both Whisper-Turbo and Whisper-Turbo Fine-Tuned.

2.4 Datasets

We used three publicly available Urdu speech datasets for training and benchmark evaluation, as

[1] https://huggingface.co/awaaz-se-alfaaz/whisper-turbo-v3-urdu

Dataset	Source	Use	Duration (in mins.)	Speech Type
Common Voice v23	Public corpus	Train + Evaluation	1267	Short Read Speech
FLEURS	Public corpus	Train + Evaluation	515	Short Read Speech
CSaLT	Public corpus	Train + Evaluation	81	Urdu Speech Corpus
YouTube Evaluation Set	Public news clips	Evaluation only	40	Two-speaker conversation

Table 1: Datasets used in training and evaluation.

shown in Table 1 : Mozilla Common Voice v23 , FLEURS , and CSaLT. Together, these datasets provide diverse speakers and recording conditions, but most utterances are short, relatively clean, and single-speaker, making them different from real-world conversational Urdu.

To evaluate real-world performance under unconstrained conditions, we created a multi-speaker evaluation set comprising 30 independent YouTube news clips totaling approximately 40 minutes of audio. Prioritizing 30 distinct clips ensured coverage across diverse speaker pairs, topics, and acoustic environments. At an average duration of roughly 80 seconds, these clips are longer than standard benchmark utterances while remaining representative of short-form news broadcasts. Expanding the dataset with longer videos would have introduced different interaction patterns and altered the scope of the evaluation; we therefore treat this dataset as a controlled real-world stress test for multi-speaker Urdu ASR.

The clips feature interactive discussions between news anchors and on-site reporters, containing natural interruptions, regional accents, background noise, and extended contextual segments. This dataset was not used during fine-tuning, and no audio clips were directly redistributed due to copyright constraints. Instead, we release the corresponding YouTube Evaluation Set as video URLs, reference transcripts, and our best-performing pipeline outputs on Hugging Face to facilitate future research.[2]

Reference transcripts for the YouTube evaluation set were created through a human-in-the-loop annotation process. Audio was first extracted from each YouTube clip via an ffmpeg-based Python pipeline, and initial transcriptions were generated using Whisper large-v3. One author (a native Urdu speaker) then listened to the audio and corrected the machine-generated transcript. A second author (also a native Urdu speaker) independently reviewed each corrected transcript while listening to the same audio, verifying and further correcting any remaining errors. Both authors followed a shared correction guideline emphasizing verbatim transcription of spoken content over normalization. Spoken content (disfluencies, code-switched words, and speaker turns) was retained in the final references, and no forced alignment was performed.

2.5 Dataset Mixing and Splits

The three public datasets were standardized into a common metadata format, including audio paths, normalized transcripts, a speaker information field, and duration information. Speaker information was included in the metadata schema but was not systematically verified or used in this work. Each dataset was independently shuffled and split into train, validation, and test subsets using an 80/10/10 ratio with a fixed random seed. The resulting train, validation, and test subsets were then pooled across datasets to form the final training, validation, and test sets used for fine-tuning and evaluation. This produced a pooled training set of 16,491 samples, comprising 13,975 from Common Voice (84.8%), 2,140 from FLEURS (13.0%), and 376 from CSaLT (2.3%). No additional balancing or weighting was applied across datasets; the pooled training set therefore reflects each dataset's natural size, with Common Voice dominating in absolute sample count.

2.6 Text and Audio Pre-processing

Experiments were run using the same pipeline for all transcripts. This encompassed Unicode Urdu normalization, removal of punctuation/symbols, standardization of Arabic/Urdu characters into various forms, number normalization, and cleanup of whitespace. Lexical content was not normalized in any other way and stemming was not applied, since this would have been detrimental to the evaluation of the ASR. All audio files were converted to mono 16 kHz, 16-bit PCM format. Leading and trailing silence were removed, and natural pauses within speech were retained. The extraction of the log-Mel spectrograms was performed as per Whisper standard.

[2] https://huggingface.co/datasets/awaaz-se-alfaaz/YouTube-Evaluation-Set

Prompt

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:

Table 2: SLM Prompt

2.7 Fine-Tuning Setup

Fine-tuning was done with AdamW with a learning rate in the range of 1e-4 to 1e-6, a batch size of 8, and epochs varied between 8 and 12. To improve the efficiency of training, the algorithm used mixed-precision training, the maximum label length was set to 128–256 tokens, and a learning-rate scheduler was employed to prevent overfitting. For LoRA, rank 32, alpha 64, and dropout 0.05 were applied to the query, key, value, and output projection layers. We did not do a full hyperparameter sweep because of compute limitations; instead, a few stability runs were performed with varying learning rate, number of epochs, and LoRA rank that were not overfitting and were not diverging on validation WER.

2.8 Benchmark Evaluation

The data used for benchmark evaluation was the pooled dataset described in 2.4: each of the three datasets (Common Voice v23, FLEURS, CSaLT) was independently split into train, validation, and test subsets using an 80/10/10 ratio, and the resulting subsets were pooled across datasets. The held-out test set was used for evaluation, with the same preprocessing, normalization, and scoring applied throughout. Results are reported for Turbo in both zero-shot and LoRA fine-tuned settings, measured using Word Error Rate (WER).

In general, we ran 30+ fine-tuning experiments with different learning rates, epoch sizes, and LoRA parameters, and selected the configuration with the best WER on the pooled validation set across all three datasets.

2.9 SLM-Based Generative Error Correction

When it comes to long-form YouTube transcriptions, we looked at Urdu-capable SLMs as text-to-text post-processing models. The SLMs were only provided with transcripts generated by the ASR and not the original audio or gold reference transcripts. This correction step was tested for both Whisper-Turbo and Whisper-Turbo Fine-Tuned, to study the effect of SLM post-processing.

All SLMs were prompted with the same conservative correction instruction: replace only incorrectly transcribed Urdu words, avoid punctuation restoration, do not add or reorder words, and preserve the original spoken style. This constrained the models to minimal ASR-error correction rather than fluent rewriting. Because the prompt restricts the SLM to replacing individual words without insertion, deletion, or reordering, this correction step is inherently limited to fixing substitution errors; insertion and deletion errors introduced by the ASR system fall outside its scope. The corrected outputs were then passed through the same normalization and jiwer-based[3] scoring pipeline used for all other experiments.

2.10 Evaluation Protocol and Metrics

All ASR and post-processing outputs were evaluated using the same normalization and scoring pipeline described in Section 2.6. The SLM correction stage was guided by a single fixed prompt, arrived at through iterative experimentation on the 30-clip YouTube evaluation set. Prompt variants were independently trialed by multiple team members across the different SLMs, with each variant assessed on its resulting WER; candidate prompts were then discussed collectively, and a single final prompt was adopted for use across all three SLMs. Each SLM was queried using its provider-recommended prompting/chat-template configura-

[3] https://github.com/jitsi/jiwer

Model / Dataset	CSaLT	Common Voice23	Fleurs
Whisper-Turbo	27.51	37.23	24.07
Whisper-Turbo Fine-Tuned	20.77	25.78	16.86

Table 3: Benchmark WER (%) for Whisper-Turbo using fine-tuning.

tion, ensuring the correction prompt was presented to every model in its intended input format.

The final prompt, shown in Table 2, was selected as it most reliably elicited minimal character- and word-level corrections without rephrasing, reordering, stylistic rewriting, or punctuation restoration. SLM correction was evaluated on both Whisper-Turbo and Whisper-Turbo Fine-Tuned outputs to separate the effect of post-processing from the effect of fine-tuning. Performance was measured using the jiwer library. WER was used as the primary metric across benchmark and YouTube evaluations.

3 Results and Findings

This section reports WER results for Whisper Turbo under zero-shot, LoRA fine-tuned settings, and SLM-based post-processing results.

3.1 Benchmark Evaluations

Table 3 summarizes WER on the three benchmark corpora for Whisper-Turbo, before and after LoRA fine-tuning. Fine-tuning yields a consistent reduction in WER across all three datasets: from 27.51% to 20.77% on CSaLT (a 24.5% relative reduction), from 37.23% to 25.78% on Common Voice v23 (a 30.8% relative reduction), and from 24.07% to 16.86% on FLEURS (a 29.9% relative reduction). In preliminary experiments using simple proportional mixing rather than stratified per-dataset splitting, we observed less consistent gains across the three corpora. Adopting stratified splitting, in which each dataset is independently divided into train, validation, and test subsets before pooling (Section 2.4), produced the uniform improvement reported above and was retained as our final experimental design.

We attribute this improvement in part to the acoustic diversity of Common Voice v23, which, despite being the noisiest of the three corpora, constitutes the majority of the pooled training set (84.8%, Section 2.4) under stratified sampling. Rather than degrading fine-tuning quality, this diversity appears to have improved the model's robustness to acoustic variation, contributing to the largest relative gain observed on Common Voice v23 itself (30.8%) while also supporting strong transfer to FLEURS (29.9%) and CSaLT (24.5%), which contributed far smaller shares of the training data (13.0% and 2.3%, respectively).

Taken together, these results indicate that LoRA fine-tuning meaningfully improves Whisper-Turbo's Urdu transcription accuracy across benchmark conditions of varying acoustic diversity and dataset scale, with exposure to Common Voice's noisier, more variable speech plausibly contributing to the model's improved generalization across all three corpora.

Model	WER (%)
Whisper-Turbo	25.33
Whisper-Turbo Fine-Tuned	24.20

Table 4: WER (%) on YouTube evaluation set.

3.2 Real-World Multi-Speaker Evaluation

To assess performance under noisy, unconstrained conditions, we evaluated our models on the curated YouTube news dataset (Section 2.4). As expected for real-world conversational audio, per-clip performance exhibits greater variability than on the standard benchmark corpora. Results fluctuate noticeably across clips, particularly those featuring strong regional accents, outdoor background noise, or colloquial speech.

Table 4 summarizes the evaluation results on this out-of-domain set. Whisper-Turbo achieves a baseline WER of 25.33%, whereas fine-tuning reduces WER to 24.20% (a relative reduction of 4.46%). However, as confirmed by our statistical significance analysis (Table 6), this 1.13% absolute gain is not statistically significant ($p = 0.8870$, 95% CI $[-4.38\%, +1.40\%]$). This confirms that acoustic fine-tuning alone fails to yield reliable improvements in unscripted, multi-speaker environments—directly motivating the need for higher-level semantic post-processing.

This relative improvement is considerably smaller than the 24.5–30.8% relative reductions observed on the benchmark corpora in Section 3.1.

SLM Model	Whisper-Turbo	Whisper-Turbo Fine-Tuned
4-bit Quantized Models		
Gemma3-12B	21.93	21.75
Qwen3-14B	23.58	22.39
Full-Precision Models		
Tiny-Aya-Fire	28.81	24.64
Qwen3-4B	30.57	29.41

Table 5: SLM post-processing Word Error Rates (WER %) across baseline and fine-tuned pipelines on the YouTube evaluation set.

	Wilcoxon Signed-Rank Test			Paired Bootstrap Resampling ($B = 10,000$)		
Comparison	**Stat (W)**	**p-value**	**Sig.**	**Mean ΔWER**	**95% CI**	**Sig.**
Turbo FT vs. Turbo	225.0	0.8870	n.s.	-1.00%	$[-4.38\%, +1.40\%]$	n.s.
Turbo+Gemma3-12B vs. Turbo	5.0	1.86×10^{-8}	***	-3.30%	$[-4.26\%, -2.41\%]$	***
Turbo FT+Gemma3-12B vs. Turbo FT	32.0	5.14×10^{-6}	***	-2.45%	$[-3.37\%, -1.44\%]$	***
Turbo FT+Qwen3-14B vs. Turbo FT	33.0	6.62×10^{-5}	***	-1.81%	$[-2.46\%, -1.12\%]$	***
Turbo FT+Qwen3-14B vs. Turbo FT+Gemma3-12B	105.0	0.0440	*	$+0.64\%$	$[+0.05\%, +1.25\%]$	*

Table 6: Statistical significance analysis across YouTube evaluation set ($N = 30$ clips). Significance levels: * $p < 0.05$, *** $p < 0.001$, n.s. = not significant.

Given the acoustic and structural differences between the fine-tuning data and the YouTube set outlined in Section 2.4, this gap suggests that fine-tuning primarily adapts the model to domain-specific training distribution rather than conferring general robustness to unscripted conversational audio. Consequently, benchmark gains do not transfer directly to noisy, real-world conditions, underscoring the necessity of evaluation sets like ours when assessing deployment readiness for Urdu ASR systems.

3.3 SLM-Based Post-Processing

To address these residual errors, we evaluate SLMs as text-to-text post-processors. Because long-form YouTube clips provide richer cross-sentence context than short benchmark utterances, they are well-suited for post-ASR error recovery while preserving the original spoken style. The SLMs were prompted conservatively to avoid adding punctuation, reordering words, rewriting sentences, or converting conversational Urdu into overly formal text.

The central challenge lies in balancing error correction with verbatim preservation. Overly aggressive correction risks normalizing colloquialisms, regional expressions, or code-switched terms into formal Urdu, ultimately worsening WER. This trade-off is particularly pronounced in clips featuring strong regional accents, informal vocabulary, or non-standard word spacing.

SLM post-processing yields mixed results depending on the corrector used, but significantly outperforms ASR fine-tuning alone (Table 6). Applying Gemma3-12B (4-bit) reduces WER to 21.93% on Whisper-Turbo (a 13.4% relative reduction; $p = 1.86 \times 10^{-8}$) and to 21.75% on Whisper-Turbo Fine-Tuned (a 10.1% relative reduction; $p = 5.14 \times 10^{-6}$), making it the strongest correction model overall. Qwen3-14B (4-bit) similarly yields statistically significant improvements across both pipelines, reducing WER to 23.58% on Whisper-Turbo (a 6.9% relative reduction) and to 22.39% on Whisper-Turbo Fine-Tuned (a 7.5% relative reduction; $p = 6.62 \times 10^{-5}$). Directly comparing the two top correctors confirms that Gemma3-12B provides a statistically significant advantage over Qwen3-14B ($p = 0.0440$).

In contrast, the two smaller full-precision models degraded performance. Tiny-Aya-Fire increases WER to 28.81% on Whisper-Turbo (a 13.7% relative increase) and to 24.64% on Whisper-Turbo Fine-Tuned (a 1.8% relative increase), while Qwen3-4B increases WER to 30.57% on Whisper-Turbo (20.7% relative increase) and to 29.41% on Whisper-Turbo Fine-Tuned (21.5% relative increase). Notably, both full-precision models degrade the baseline pipeline more severely than the fine-tuned pipeline, suggesting that a stronger starting transcript is somewhat more resilient to unhelpful corrections than a noisier one.

Model / Stage	Transcription Text (WER)
Raw ASR	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 آجاتی ہے جیسے یہ فلک کرتے ہیں بال کو کنگز کے پاس دوینگ اچھے فاس بولرز بھی ہیں عباس آفریڈی 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 آفریڈی نئے بال پہ یورکر کروا سکتے ہیں بہت اچھے دو یینگ 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
GEC: Gemma3-12B (4-bit)	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 عرشد اقبال عباس آفریدی اور محمد حارث لمبے عرصے تک کراچی کنگز کے لئے کھیل سکتے ہیں کنگز کے پاس یہ ینگ پلیز ہیں جو آپ کی توجہوں کا مرکز بن سکتے ہیں آگئے چل قاسم اکرم گوڈ لکنگ بیٹسمن ہے لیک سائیڈ پہ جو ان کی بیٹنگ ہے وہ ویراٹ کولی کی آپ کو یاد آجاتی ہے جیسے یہ فلک کرتے ہیں بال کو کنگز کے پاس دوینگ اچھے فاس بولرز بھی ہیں عباس آفریدی اور عرشد اقبال لمبے کٹ کے ہیں سپیڈ بڑی اچھی یا ہارڈ لینکھ 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 یینگ 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 حارث بہت اچھی پوٹینچل ہے یہ بھی ٹیمپرومنٹلی سٹرانگ ہے اور کالیٹیز بے انتہاں ہیں

Table 7: Qualitative comparison of Raw ASR vs. Post-SLM corrected Urdu transcripts. Highlighted text indicates residual recognition, orthographic, or spacing errors.

Comparing correction gains across the Whisper-Turbo baseline and fine-tuned pipelines also suggests differing relationships between SLM correction and fine-tuning. For Gemma3-12B, the relative gain from correction is larger on the Whisper-Turbo baseline (13.4%) than on the already fine-tuned pipeline (10.1%), consistent with SLM correction and fine-tuning addressing at least partially overlapping error patterns. For Qwen3-14B, the relative gain is similar across both pipelines (6.9% vs. 7.5%), suggesting a more complementary relationship between correction and fine-tuning for this model. The best overall WER, 21.75%, is achieved by combining fine-tuning with Gemma3-12B correction.

A qualitative examination of the post-processed outputs reveals distinct failure modes that explain the degradation observed in smaller models. Despite strict prompts prohibiting formatting changes, Tiny-Aya-Fire consistently inserted unprompted Urdu punctuation (full stops '۔' and commas), which artificially inflated edit distances against unpunctuated reference transcripts. Qwen3-4B exhibited even more severe instruction-following breakdowns: on certain clips, it entered repetitive generation loops (reproducing the transcript up to three times consecutively), while on another, it translated the Urdu transcript into English. Because these

Error Source	Observed Effect
Regional accents and colloquial pronunciation	Increased substitutions and non-standard lexical forms
Outdoor noise and rapid speaker alternation	Missed or confused words during turn changes
Urdu orthographic variation	Semantically correct outputs were penalized due to spacing or alternate spellings
SLM over-correction	Conversational Urdu sometimes normalized into overly formal wording

Table 8: Error Analysis.

catastrophic failure modes yielded clip-level WERs approaching 100%, they disproportionately inflated Qwen3-4B's average error rate. In contrast, Table 7 demonstrates a successful case where Gemma3-12B (4-bit) corrects several recognition and orthographic errors while preserving the underlying spoken content.

3.4 Error Analysis

A closer examination of the YouTube evaluation showed that a few common conditions repeatedly resulted in high WER.

Table 8 summarizes the main error sources observed in the YouTube evaluation set and their effect on transcription quality. WER increased with stronger regional accents, outdoor or noisy recording conditions, faster speaker alternation, and more colloquial speech. These conditions are not well represented in the benchmark datasets, which tend to be shorter, cleaner, and structurally simpler.

A further source of measured error was Urdu's orthographic variation itself, rather than genuine recognition failure: outputs that were semantically correct were still penalized by WER when they differed from the reference in spacing conventions, compound-word segmentation, or alternate valid spellings. SLM-based cleanup was more effective on clips that were more lexically dense and contextually continuous, and less effective on clips that were sparser or more fragmented.

GEC Model	Per Min (s)
Tiny-Aya-Fire	5.5
Qwen3-4B	10.5
Gemma3-12B (4-bit)	7.5
Qwen3-14B (4-bit)	12.5

Table 9: Inference latency for the SLM GEC stage.

3.5 Deployment Considerations and Runtime Trade-offs

Beyond transcription accuracy, practical deployment of Urdu ASR systems depends on processing speed. We therefore report latency, measured in seconds per minute of audio processed, for each pipeline configuration alongside the corresponding WER results from Section 3.3, to characterize the trade-off between correction quality and computational cost.

ASR inference alone is fast, with fine-tuning introducing no latency overhead while improving accuracy. However, adding SLM-based correction increases processing time in exchange for varying accuracy gains. Table 9 details the latency added by each GEC model in seconds per minute of audio processed.

Among the evaluated correctors, Gemma3-12B (4-bit) offers the strongest performance-to-latency trade-off: as shown in Table 9, it requires only 7.5 seconds per minute of audio to achieve the best overall WER of 21.75%. Notably, despite having three times as many parameters, Gemma3-12B (4-bit) processes audio faster than Qwen3-4B (7.5s vs. 10.5s per minute) while yielding significantly better accuracy gains. Qwen3-14B requires the highest runtime overhead at 12.5 seconds per minute of audio, whereas Tiny-Aya-Fire is the fastest corrector at 5.5 seconds per minute, though at the expense of severe accuracy degradation.

As noted in Section 2.3, we additionally tested Gemma3-12B at 8-bit quantization; while WER improved marginally (21.68% vs. 21.75% at 4-bit), inference latency increased substantially to nearly five times that of the 4-bit variant due to hardware limitations. This confirms 4-bit quantization as the more practical configuration for real-time and resource-constrained deployment.

4 Conclusion

This work establishes updated baselines for Urdu ASR using Whisper-Turbo, evaluated across curated benchmark corpora and a real-world YouTube

evaluation set. LoRA fine-tuning yields consistent WER reductions across all three benchmark datasets (24.5–30.8% relative), driven largely by Common Voice v23's acoustic diversity, but this gain transfers less strongly to real-world YouTube audio (4.46% relative), underscoring a persistent benchmark-to-deployment gap.

SLM-based GEC offers the largest further improvement, but its benefit is highly model-dependent: Gemma3-12B (4-bit) achieves the best overall result (21.75% WER, a 10.1% relative reduction over fine-tuned Whisper-Turbo), whereas Tiny-Aya-Fire and Qwen3-4B degrade performance. Overall, combining targeted fine-tuning with a carefully selected, lightweight SLM corrector offers a practical path toward more robust Urdu ASR for real-world, multi-speaker audio. To support reproducible research and benchmark reuse, we make our fine-tuned Whisper-Turbo checkpoint and the 30-clip YouTube evaluation corpus publicly available on Hugging Face.

In future work, we plan to explore decoder-level hyperparameter optimization (such as beam search, N-best reranking, and temperature scaling) alongside fine-tuning and SLM correction, as our preliminary tests showed promising WER reductions on FLEURS without modifying model weights. Additionally, future efforts will focus on disentangling model scale from quantization effects, developing domain-adapted correction models to mitigate over-formalization of colloquial speech, and expanding the YouTube evaluation set for community reuse.

5 Ethics

All the training sets used in this work were obtained from public resources: Mozilla Common Voice v23, FLEURS and CSaLT, all of which are publicly available, and used within their respective terms. YouTube Clips were not used to train, redistribute, or release, respecting copyright and platform policies. No personally identifiable or private data was collected.

The only post-processing done in the SLM system was minimal correction, but language models might have favored standard Urdu over other Urdu variants. We thus stress the need for careful deployment of Urdu ASR systems, taking into consideration privacy, fairness, and dialects.

Acknowledgments

The authors would like to thank the anonymous reviewers for their detailed feedback, which significantly improved the quality of this paper.

References

Afsheen, Saad Irtza, Mahwish Farooq, and Sarmad Hussain. 2014. Accent classification among punjabi, urdu, pashto, saraiki and sindhi accents of urdu language. In *Proceedings of the Conference on Language and Technology (CLT 2014)*. Center for Language Engineering (CLE), UET Lahore.

Harsh Ahlawat, Naveen Aggarwal, and Deepti Gupta. 2025. Automatic speech recognition: A survey of deep learning techniques and approaches. *International Journal of Cognitive Computing in Engineering*, 6:201–237.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 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).

Samee Arif, Aamina Jamal Khan, Mustafa Abbas, Agha Ali Raza, and Awais Athar. 2025a. CSaLT-voice. Hugging Face.

Samee Arif, Aamina Jamal Khan, Mustafa Abbas, Agha Ali Raza, and Awais Athar. 2025b. 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.

Cohere Labs. 2024. Tiny aya fire. Hugging Face.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. FLEURS: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Rao Ma, Mengjie Qian, Potsawee Manakul, Mark Gales, and Kate Knill. 2023. Can generative large language models perform ASR error correction? ArXiv preprint arXiv:2307.04172.

Qwen Team. 2025. Qwen3 technical report. ArXiv preprint arXiv:2505.09388.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International*

Conference on Machine Learning, volume 202, pages 28492–28518, Honolulu, Hawaii, USA. PMLR.

Najm Ul Sehar, Ayesha Khalid, Farah Adeeba, and Sarmad Hussain. 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.

Natsuo Yamashita, Masaaki Yamamoto, Hiroaki Kokubo, and Yohei Kawaguchi. 2025. LLM-based generative error correction for rare words with synthetic data and phonetic context. In *Proceedings of Interspeech 2025*, pages 3653–3657, Rotterdam, The Netherlands. International Speech Communication Association (ISCA).

Arturs Znotins, Normunds Gruzitis, and Roberts Dargis. 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.