

Improving Urdu ASR with Fine-Tuned Small Language Models for Post-ASR Error Correction

Sajjad Haider, Solat J. Sheikh, Zuha Aqib, Farah Inayat, Zehra Ahmed

Artificial Intelligence Lab

Department of Computer Science

Institute of Business Administration (IBA), Karachi, Pakistan

{sahaidar, sjsheikh}@iba.edu.pk

{z.aqib.26106, f.inayat.26912, z.ahmed.26965}@khi.iba.edu.pk

Abstract

Current multilingual automatic speech recognition (ASR) models have substantially improved speech recognition across many languages, yet recognition quality remains limited for low-resource languages such as Urdu. A common approach to improve recognition quality is to fine-tune the ASR model itself, but this requires labelled speech data and re-training the recognizer. A recently emerging alternative is Generative Error Correction (GEC), in which a language model post-processes ASR output to correct transcription errors while leaving the underlying speech recognizer unchanged. This paper investigates this approach for Urdu by keeping Whisper-Turbo frozen and fine-tuning small language models to correct its transcription errors. To address data scarcity in Urdu, we construct a pseudo-gold training dataset using a strong teacher model. Fine-tuning 3.5–4B parameter language models on this data achieves a 33% relative WER reduction (from 11.74% to 7.81%), outperforming zero-shot baselines as well as significantly larger 12–14B zero-shot models. These results show that combining a frozen recognizer with a small fine-tuned corrector provides a practical, resource-efficient framework for Urdu ASR improvement without relying on ASR retraining or substantially larger models.

1 Introduction

Over the past few years, multilingual automatic speech recognition (ASR) systems, such as Whisper (Radford et al., 2023), Qwen3-ASR (Shi et al., 2026), and NVIDIA Nemotron 3.5 ASR (Noroozi et al., 2024), have expanded speech processing across global languages. However, despite being spoken by approximately 246 million people worldwide,¹ Urdu remains underrepresented in

speech technologies and continues to exhibit relatively high error rates in multilingual ASR benchmarks. Among these models, Whisper is currently the only one with native support for Urdu, making improvements to its transcription output particularly valuable.

The standard approach to addressing high error rates is fine-tuning the speech recognizer on additional language-specific speech data. While a few studies have attempted to fine-tune Whisper variants for Urdu, they report modest improvements over the baseline (Arif et al., 2025; Sehar et al., 2025; Ahmed et al., 2026). Furthermore, this approach requires paired audio-text data, which remain scarce for Urdu, and requires costly retraining on speech pipelines whenever domain adaptation is needed.

An alternative is to leave the ASR model unchanged and correct its output with a language model. This approach, known as Generative Error Correction (GEC) (Yang et al., 2023; Ma et al., 2023), treats error correction as a text-to-text generation task: the speech recognizer produces an initial transcript, and a language model rewrites it to correct substitution and spelling errors while preserving the underlying meaning. While prior work has demonstrated the effectiveness of GEC across a variety of settings, most existing approaches rely on zero-shot prompting of large frontier language models or benchmark development, rather than fine-tuning small open-weight language models for low-resource languages.

As a result, it remains unclear whether the reported improvements primarily stem from the scale of frontier language models or from task-specific fine-tuning, since GEC for low-resource languages has rarely been studied using fine-tuned correction models. A major practical challenge to investigating this question is the lack of human-annotated correction corpora needed for supervised training. These challenges are particularly

¹<https://www.ethnologue.com/insights/ethnologue200/> (accessed July 26, 2026)

acute for Urdu, where both labelled speech and human-corrected transcripts are scarce.

To address these challenges, this paper investigates a Generative Error Correction framework based on fine-tuning small language models for Urdu. The underlying Whisper-Turbo² speech recognizer remains frozen, while a small language model is trained to correct its output. To overcome the lack of human-annotated GEC corpora, we construct a pseudo-gold³ training dataset using a teacher-student pipeline. Short Urdu speech segments from diverse YouTube sources are first transcribed with Whisper-Turbo. After filtering out low-confidence transcriptions, Google’s Gemini 3 Flash Preview⁴ is used as the teacher model to generate refined transcripts that serve as pseudo-gold targets. These paired transcripts are then used to fine-tune two small language models: Tiny-Aya-Fire (3.5B) (Salamanca et al., 2026) and Qwen3-4B-Instruct (Team, 2025).

We evaluate the proposed approach against three distinct baselines: uncorrected Whisper-Turbo, zero-shot small language models (Tiny-Aya-Fire and Qwen3-4B-Instruct), and zero-shot substantially larger language models, specifically Gemma3-12B⁵ and Qwen3-14B⁶. Fine-tuning small models on pseudo-gold data yields consistent improvements, reducing WER from 11.74% to 7.81% with Qwen3-4B-Instruct and 7.83% with Tiny-Aya-Fire, corresponding to a relative reduction of approximately 33%. Section 4 reports these results in full, including non-parametric significance tests comparing fine-tuned small models against substantially larger zero-shot baselines, and an analysis of how training set size affects downstream performance.

2 Related Work

Generative Error Correction (GEC) is an approach to improve Automatic Speech Recognition (ASR) performance by using pretrained Language Models

(LMs) to post-process the ASR output text, without changing the speech recognizer itself. Early GEC work established that large language models can effectively correct ASR 1-best or N -best outputs (Yang et al., 2023; Ma et al., 2023). These studies mostly used English benchmarks and showed that zero-shot or few-shot in-context learning (ICL) with closed frontier LLMs can fix many grammatical and semantic errors. More recent work has extended GEC to other languages, such as Japanese (Ko et al., 2026).

While most of this work relies on zero-shot or few-shot prompting, a smaller body of research has investigated task-specific fine-tuning of language models for ASR error correction. Yang et al. (2023) showed that LoRA fine-tuning an open LLM gives large gains over ICL alone on English data. Li et al. (2024) fine-tuned a multilingual LLM to correct 1-best ASR hypotheses across more than 100 languages, demonstrating cross-lingual knowledge transfer for lower-resource languages sharing the same writing system. Similarly, Yamashita et al. (2025) fine-tuned a small correction model on machine-generated transcription pairs.

Recent work has also explored prompting-based GEC beyond English. Monika and Passricha (2026) proposed a prompting framework that combines In-Context Learning (ICL) and Task-Activating Prompting (TAP) for Hindi and Punjabi ASR correction without fine-tuning. Similarly, Hao et al. (2026) conducted a contamination-aware evaluation of LLM-based GEC for West Frisian. Despite encouraging results, these approaches relied on large frontier LLMs and zero-shot or few-shot prompting, leading to high inference cost and limited suitability for local deployment.

Closely related to our setting, Ahad et al. (2026) construct annotated corpora for Urdu grammatical error correction, though their focus is text-based GEC rather than correction of ASR output. For Hindi, a linguistically related low-resource language, Kumar et al. (2026) compare fine-tuned language models against large language models for post-ASR correction and similarly find that scale alone does not determine correction quality. Ghosh et al. (2025) improve generative ASR error correction by augmenting training data with synthetic examples and retrieval-augmented context, a strategy orthogonal to the teacher-generated pseudo-gold approach used here. Kumar et al. (2024) address a related but narrower problem, us-

²A faster variant of Whisper Large-v3. See: <https://github.com/openai/whisper/discussions/2363> (accessed July 26, 2026).

³We use “pseudo-gold” and “silver-standard” interchangeably to denote automatically generated labels that approximate, but are not guaranteed to match, gold-standard human annotations.

⁴<https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview>. Accessed July 26, 2026.

⁵<https://huggingface.co/google/gemma-3-12b-it> (accessed July 26, 2026).

⁶<https://huggingface.co/Qwen/Qwen3-14B> (accessed July 26, 2026).

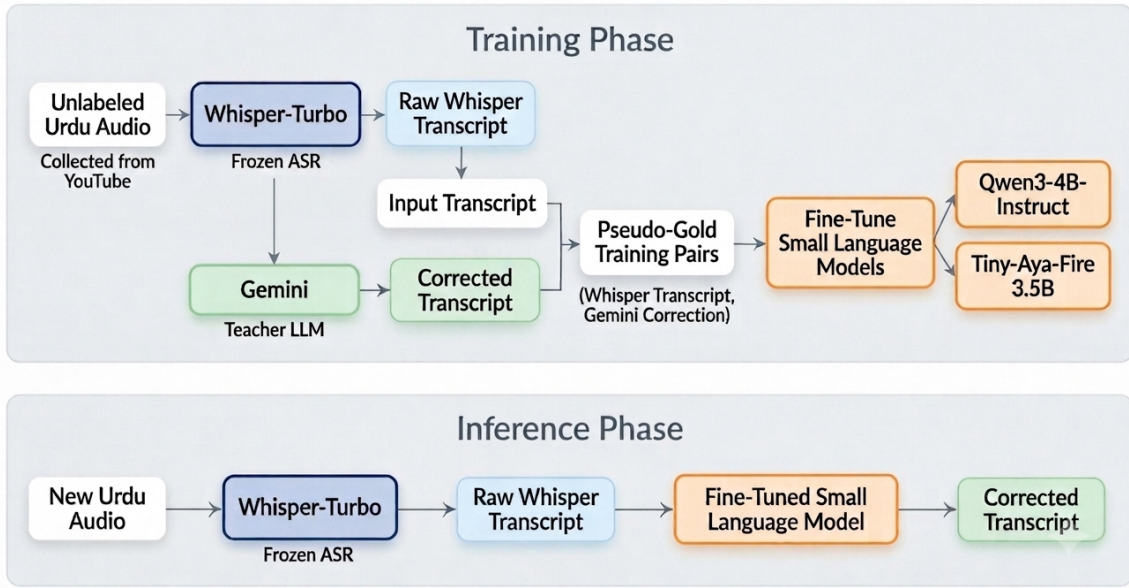


Figure 1: Overview of the Generative Error Correction framework showing the pseudo-gold training pair generation phase (top) and the fine-tuned inference pipeline (bottom).

ing dictionary-augmented prompting to improve cross-lingual proper noun recognition in ASR output, complementing the general-purpose correction studied in this work.

Taken together, the existing literature points to a clear gap. Fine-tuning has consistently improved GEC performance when suitable training data are available, whereas work on lower-resource languages has relied almost exclusively on zero-shot or few-shot prompting with large frontier LLMs. A major obstacle to extending fine-tuning to lower-resource languages is the lack of parallel training pairs consisting of ASR outputs and reference transcripts. To address this limitation, we generate pseudo-gold transcription pairs for Urdu and use them to fine-tune small language models while keeping the Whisper-Turbo ASR model frozen.

3 Methodology

Figure 1 provides an overview of the teacher-student pipeline used to automatically generate training pairs from real-world speech, encompassing dataset curation, model fine-tuning, and evaluation.

3.1 Pseudo-Gold Dataset Construction

A central contribution of this work is a pseudo-gold dataset construction pipeline that enables supervised fine-tuning for Urdu post-ASR correc-

tion despite the scarcity of manually corrected Urdu transcripts, using a teacher-student approach, detailed below, to automatically generate training pairs at scale. Approximately 48 hours of Urdu audio were collected from publicly available YouTube videos covering a variety of speaking styles, including podcasts, talk shows, news broadcasts, reporter diaries, and reporter-anchor interactions.

The recordings were first segmented into clips of approximately 10–30 seconds using silence-aware segmentation. For each recording, the first sufficiently long pause occurring after the minimum segment duration was selected as the split point. If no suitable pause was detected, the segment was cut at the maximum duration. This strategy produced naturally bounded utterances while avoiding unnecessary fragmentation of continuous speech, yielding approximately 6,400 speech segments.

Each segment was then transcribed using Whisper-Turbo. Before generating training labels, the resulting transcriptions were filtered to remove unreliable ASR outputs. Samples were discarded if they fell outside predefined word-count limits (as short clips frequently suffered from under-recognition while overly long sequences were prone to repeated n -gram hallucinations), contained insufficient Urdu script, or showed unusually low lexical diversity. After filtering, 5,225 seg-

Model	Average (%)	Median (%)	Max (%)
Whisper-Turbo (Baseline)	11.74	10.59	35.06
<i>Zero-shot Correctors</i>			
Tiny-Aya-Fire (3.5B)	10.41	8.97	36.36
Qwen3-4B-Instruct	9.89	8.84	28.57
Qwen3-14B	9.04	8.07	28.57
Gemma-12B	8.32	7.32	24.56
<i>Fine-tuned Qwen3-4B</i>			
Qwen_exp_1600	9.03	7.94	49.48
Qwen_exp_3200	8.21	6.88	33.33
Qwen_exp_4800	7.81	6.67	28.89
<i>Fine-tuned Tiny-Aya-Fire (3.5B)</i>			
Tinyaya_exp_1600	8.72	7.56	48.28
Tinyaya_exp_3200	8.26	6.84	38.57
Tinyaya_exp_4800	7.83	6.38	40.58

Table 1: WER (%) across all evaluated configurations on the 425-pair held-out test set.

ments (approximately 40 hours of audio) were retained for pseudo-gold generation.

These filtered Whisper transcripts were then corrected using Gemini 3 Flash Preview⁷ using the prompt shown in Appendix A.1. The teacher model was instructed to correct only transcription errors while preserving the original meaning and to return only the corrected Urdu transcript. The resulting Whisper–Gemini transcription pairs formed the pseudo-gold corpus used throughout this study.

To assess the quality of the automatically generated labels, 100 transcription pairs were randomly selected and manually checked against the source audio by two of the authors, both native Urdu speakers. Gemini’s corrections achieved an average WER of less than 2.5% based on this manual check, indicating that the pseudo-gold labels were sufficiently accurate for training.

The final corpus contained 5,225 Whisper–Gemini transcription pairs. A held-out set of 425 pairs (approximately 4 hours) was reserved for evaluation, while the remaining data were used for training. To study the effect of training data size, three progressively larger training subsets containing 1,600 (12 hours), 3,200 (24 hours), and 4,800 examples (36 hours) were constructed.

3.2 Models and Fine-Tuning

Two instruction-tuned small language models were selected for post-ASR error correction: Tiny-Aya-Fire (3.5B parameters), designed for multilingual applications with an emphasis on South Asian

languages, and Qwen3-4B-Instruct, a multilingual instruction-tuned model of comparable size. Substantially larger open-weight models, such as Gemma-12B and Qwen3-14B, were evaluated in the zero-shot setting to compare small fine-tuned models against larger un-tuned baselines.

Both small language models were fine-tuned using supervised fine-tuning (SFT) with parameter-efficient Low-Rank Adaptation (LoRA). LoRA modules ($r = 16$, $\alpha = 32$, dropout = 0.05) were applied to the query, key, value, and output projection matrices of the attention layers, while the remaining model parameters remained frozen. Training was performed for three epochs using the AdamW optimizer with a learning rate of 3×10^{-4} , cosine learning-rate scheduling, an effective batch size of 16, and fp16 mixed-precision training. All experiments used a maximum sequence length of 1,024 tokens and were conducted on a single NVIDIA RTX-4080 GPU (16 GB). It should be noted that no separate development set was used for hyperparameter or prompt selection. A small number of preliminary trials with LoRA rank and batch-size configurations were run only to ensure training completed without out-of-memory errors while making full use of the available 16GB of GPU memory, and these trials were not used to tune values based on downstream correction performance; the hyperparameters above were otherwise fixed in advance.

A unified instruction template (Appendix A.2) was used for fine-tuning and zero-shot inference across Tiny-Aya-Fire and Qwen3-4B-Instruct so that performance differences reflect task-specific adaptation rather than prompt engineering. Com-

⁷The model was accessed through the OpenRouter API (<https://openrouter.ai/>).

Comparison	Wilcoxon Signed-Rank Test			Paired Bootstrap Resampling		
	Stat (W)	p-value	Sig.	Mean Δ WER	95% CI	Sig.
FT Tiny-Aya vs. Raw Whisper	6,984.0	1.07×10^{-47}	***	-0.0391	[-0.0435, -0.0347]	***
FT Qwen3-4B vs. Raw Whisper	9,306.0	1.08×10^{-41}	***	-0.0392	[-0.0439, -0.0345]	***
FT Tiny-Aya vs. ZS Tiny-Aya	6,748.5	8.05×10^{-30}	***	-0.0258	[-0.0304, -0.0212]	***
FT Qwen3-4B vs. ZS Qwen3-4B	11,113.5	1.58×10^{-25}	***	-0.0208	[-0.0247, -0.0169]	***
FT Tiny-Aya (3.5B) vs. ZS Qwen3-14B	13,739.0	1.33×10^{-14}	***	-0.0121	[-0.0157, -0.0084]	***
FT Qwen3-4B (4B) vs. ZS Gemma3-12B	20,009.5	1.82×10^{-4}	***	-0.0051	[-0.0088, -0.0013]	**
FT Qwen3-4B vs. FT Tiny-Aya	23,089.0	0.798	n.s.	-0.0002	[-0.0041, +0.0036]	n.s.

Table 2: Statistical significance analysis across per-utterance WER distributions.

pared with the simpler teacher prompt used to generate the pseudo-gold corpus (Appendix A.1), this template provides more explicit constraints, instructing the model to correct only transcription errors, preserve correct words unchanged, and return only the corrected transcript. The more detailed prompt was adopted after preliminary experiments showed that simpler instructions occasionally produced explanatory text or modified words that were already correct.

Each model was fine-tuned independently on the three training subsets (1,600, 3,200, and 4,800 examples), yielding six fine-tuned correction models. Due to GPU memory limitations, larger models like Gemma-12B and Qwen3-14B were evaluated only in the zero-shot setting. This experimental design allows us to evaluate two central questions: first, whether post-ASR correction improves upon raw Whisper-Turbo transcriptions; and second, whether fine-tuning of small open-weight models outperforms zero-shot prompting of substantially larger architectures.

4 Experiments and Results

This section evaluates the proposed post-ASR correction pipeline. It outlines the experimental setup, compares baseline and fine-tuned models across dataset scales, measures inference latency, and analyzes qualitative correction examples.

4.1 Experimental Setup

Performance was evaluated using Word Error Rate (WER), where lower values indicate better transcription quality. Evaluation was performed on the held-out test set of 425 pseudo-gold transcription pairs as described earlier. Since the reference transcripts were generated automatically using Gemini, the reported WER measures agreement with the teacher-generated corrections under a consistent evaluation protocol. In addition to transcrip-

tion accuracy, we measured the inference time of the post-ASR correction stage over the entire evaluation set to quantify its computational overhead. It is worth mentioning that we additionally computed Character Error Rate (CER) for all configurations. CER tracked WER closely across models, with the configuration achieving the lowest WER also achieving the lowest CER, so we report WER only throughout for brevity.

4.2 Overall Performance

Table 1 reports WER across all evaluated systems on the held-out test set.

Every correction model improves on the Whisper-Turbo baseline, though the magnitude varies substantially. Among zero-shot systems, Gemma3-12B performs best (8.32%), followed by Qwen3-14B (9.04%), Qwen3-4B (9.89%), and Tiny-Aya-Fire (10.41%). Larger models show stronger intrinsic correction ability even without task-specific adaptation.

Fine-tuning, however, produces substantially larger gains than zero-shot prompting alone. Qwen3-4B fine-tuned on 4,800 examples reaches the lowest WER (7.81%); Tiny-Aya-Fire reaches a nearly identical 7.83%. Relative to the 11.74% baseline, both correspond to an approximately 33% relative reduction.

Critically, both fine-tuned 3.5–4B models outperform the much larger 12–14B zero-shot models. Modest supervision outweighs a roughly $3\times$ – $4\times$ increase in parameter count in this setting. Figure 2 shows the distribution of per-utterance WER for each configuration. As training data increases, the distributions for both Qwen3-4B and Tiny-Aya-Fire shift downward, indicating consistent improvements across the test set as reported in Table 1. This suggests that the gains from fine-tuning are not limited to a small number of utterances but are consistently observed across the evaluation set.

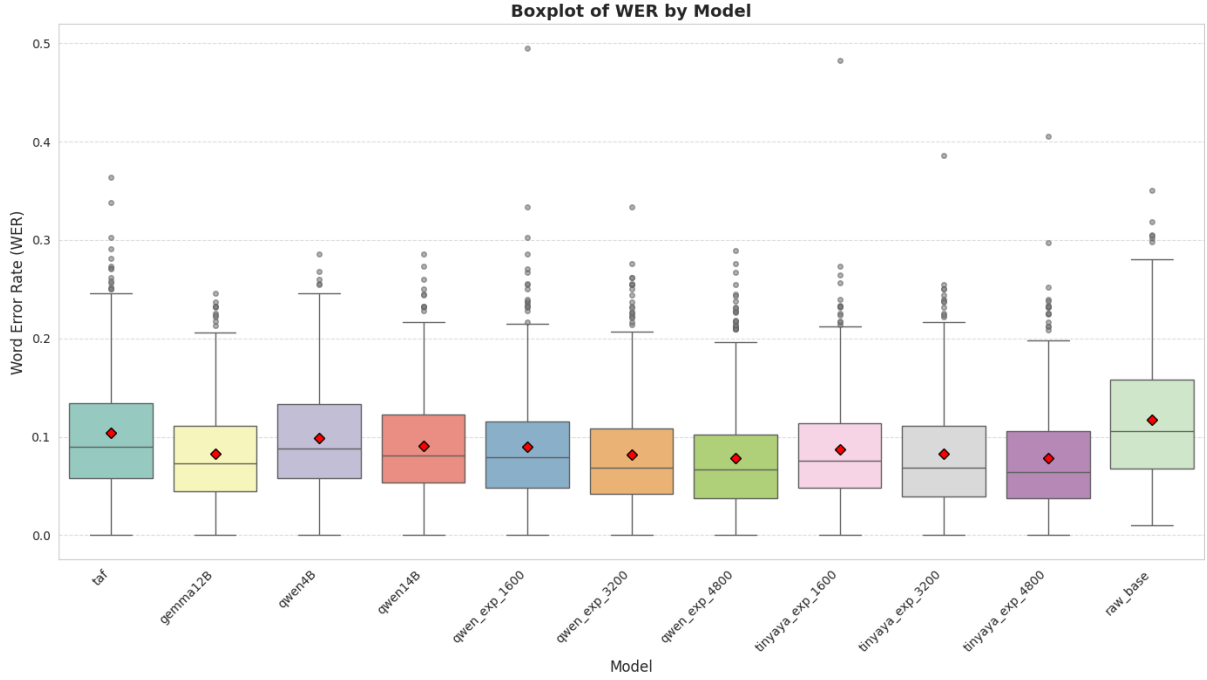


Figure 2: Distribution of per-utterance Word Error Rates (WER) across baseline and fine-tuned models.

To confirm that the observed WER reductions reflect statistically meaningful gains, we conducted both non-parametric Wilcoxon signed-rank tests and paired bootstrap resampling ($N = 10,000$ iterations) across per-utterance error distributions (Table 2).

Evaluating these statistical metrics alongside the overall WER performance in Table 1 strongly validates our key hypotheses. Fine-tuning yields a statistically significant absolute WER reduction of 3.9 percentage points over the raw Whisper-Turbo baseline (11.74% $\rightarrow$ 7.81%, representing a 33.5% relative error reduction; $p < 0.0001$, 95% CI: [-4.39%, -3.45%]). Furthermore, fine-tuning significantly outperforms zero-shot evaluation of the same architectures ($p < 0.0001$). Fine-tuned small models (3.5B to 4B) achieve statistically superior performance compared to substantially larger zero-shot baselines (12B to 14B): fine-tuned Tiny-Aya-Fire (3.5B) outperforms Qwen3-14B zero-shot by 1.21 percentage points in WER (9.04% $\rightarrow$ 7.83%, $p < 0.0001$, 95% CI: [-1.57%, -0.84%]), while fine-tuned Qwen3-4B outperforms Gemma3-12B zero-shot by 0.51 percentage points (8.32% $\rightarrow$ 7.81%, $p < 0.001$, 95% CI: [-0.88%, -0.13%]). Finally, a direct comparison between the two fine-tuned small models confirms no statistically significant difference ($p = 0.798$, 95% CI: [-0.41%, +0.36%]), demonstrating that pseudo-gold adaptation allows distinct

3.5B to 4B models to achieve equivalent high-quality error correction.

4.3 Qualitative Analysis and Error Categorization

To better understand the types of errors corrected by fine-tuning, Table 3 provides a qualitative comparison across two representative test utterances. An analysis of the output sequences highlights three main categories of errors corrected by fine-tuning:

1. **Proper Noun and Named Entity Recovery:** Acoustic models frequently struggle with domain-specific proper nouns due to vocabulary limitations or acoustic ambiguities. In Example 1, raw Whisper-Turbo phonetically misidentifies the proper name *Hashemi Rafsanjani* as *حاشمی رفزنجانی* (erroneously substituting ز for س). In Example 2, Whisper severely misrecognizes the Pakistani political figure *Bilawal Bhutto Zardari* (بلاول بوتل زرداری). While zero-shot Qwen3-4B fails to recover these entities (yielding بلاول بوتل زرداری), fine-tuning uses context to correctly restore the full name (بلاول بھٹو زرداری).
2. **Word Segmentation and Compounding:** Raw ASR transcriptions in Urdu often miss

Model / Source	Transcription Text
<i>Example 1</i>	
Whisper-Turbo (Raw)	حاشی رفرنجانی نے آپسی معاملات کو نرم کرنے اور عراق کے مسئلے پر باتچیت کے لیے ایک ڈائلاگ پروسس کا آغاز کیا اس ڈائلاگ پروسس کو سعودی عرب سے
Qwen3-4B (Zero-Shot)	حاشی رفرنجانی نے آپسی معاملات کو نرم کرنے اور عراق کے مسئلے پر باتچیت کے لیے ایک ڈائلاگ پروسس کا آغاز کیا اس ڈائلاگ پروسس کو سعودی عرب سے
Gemma3-12B (Zero-Shot)	حاشی رفرنجانی نے آپسی معاملات کو نرم کرنے اور عراق کے مسئلے پر باتچیت کے لیے ایک ڈائلاگ پروسس کا آغاز کیا اس ڈائلاگ پروسس کو سعودی عرب سے
Qwen3-4B (Fine-Tuned)	حاشی رفرنجانی نے آپسی معاملات کو نرم کرنے اور عراق کے مسئلے پر بات چیت کے لیے ایک ڈائلاگ پروسس کا آغاز کیا اس ڈائلاگ پروسس کو سعودی عرب سے
Pseudo-Gold Target	حاشی رفرنجانی نے آپسی معاملات کو نرم کرنے اور عراق کے مسئلے پر بات چیت کے لیے ایک ڈائلاگ پراسیس کا آغاز کیا۔ اس ڈائلاگ پراسیس کو سعودی عرب سے
<i>Example 2</i>	
Whisper-Turbo (Raw)	ہماری کچھ سیٹے کم کی گئی ہیں لیکن پھر بھی جی بی کے انٹرسٹ میں گلگت بلتستان کی عوام کے مفاد کے لیے ہم نے ایک اوپن ڈور پولیسی رکھی ہے بلاول پوٹر تدری
Qwen3-4B (Zero-Shot)	ہماری کچھ سیٹیں کم کی گئی ہیں لیکن پھر بھی جی بی کے اینٹرسٹ میں گلگت بلتستان کی عوام کے مفاد کے لیے ہم نے ایک اوپن ڈور پالیسی رکھی ہے بلاول پوتل تدری
Gemma3-12B (Zero-Shot)	ہماری کچھ سیٹیں کم کی گئی ہیں لیکن پھر بھی جی بی کے انٹرسٹ میں گلگت بلتستان کی عوام کے مفاد کے لیے ہم نے ایک اوپن ڈور پالیسی رکھی ہے بلاول پوتر تدری
Qwen3-4B (Fine-Tuned)	ہماری کچھ سیٹیں کم کی گئی ہیں لیکن پھر بھی جی بی کے انٹرسٹ میں گلگت بلتستان کی عوام کے مفاد کے لیے ہم نے ایک اوپن ڈور پالیسی رکھی ہے بلاول بھٹو زرداری
Pseudo-Gold Target	ہماری کچھ سیٹیں کم کی گئی ہیں لیکن پھر بھی جی بی کے انٹرسٹ میں گلگت بلتستان کی عوام کے مفاد کے لیے ہم نے ایک اوپن ڈور پالیسی رکھی ہے بلاول بھٹو زرداری

Table 3: Qualitative comparison of post-ASR error correction outputs across models. Highlighted words indicate deviations from the pseudo-gold target.

spaces between compound words. In Example 1, fine-tuning correctly inserts word boundaries into concatenated tokens such as بات چیت → بات چیت, bringing the sequence into alignment with standard Urdu orthography.

3. Loanword and Diacritic Standardization:

English loanwords in Urdu news broadcasts (e.g., *dialogue*, *policy*, *interest*) often appear with inconsistent spellings in raw ASR output (e.g., ڈائلوگ vs. ڈائلوگ). Fine-tuning consistently maps these variants onto canonical Urdu loanword representations (ڈائلوگ). Additionally, implicit morphological markers, such as plural noun endings (سیٹھ → سیٹھیں), are accurately restored.

Overall, these qualitative examples confirm that fine-tuning helps the language model learn Urdu error patterns that zero-shot prompting misses. By adapting directly to ASR output, the model effectively resolves phonetic errors, missing spaces, and inconsistent loanword spellings that cannot be fixed by acoustic signals alone.

4.4 Effect of Training Data Size

Table 1 also illustrates the effect of increasing pseudo-gold training data. Both models benefit consistently: Qwen3-4B improves from 9.03% (1,600 examples) to 8.21% (3,200) to 7.81% (4,800); Tiny-Aya-Fire follows a similar trend, from 8.72% to 8.26% to 7.83%. The rate of improvement slows as training data increases, but WER continues to decrease even at the largest training set evaluated, suggesting that additional pseudo-gold data may yield further gains.

The scale of this improvement is clearest against each model’s own zero-shot baseline: Qwen3-4B improves from 9.89% zero-shot to 7.81% fine-tuned; Tiny-Aya-Fire improves from 10.41% to 7.83%. These consistent shifts demonstrate that adaptation provides substantial error-correction capability beyond zero-shot instruction following.

4.5 Inference Latency

Table 4 reports the total inference time required by each GEC model to process the complete test set (approximately four hours of audio), along with the equivalent processing time per minute of input audio. Tiny-Aya-Fire required the least processing

time, averaging only 7.4 seconds of post-ASR correction per minute of speech. Interestingly, inference time did not scale solely with model size. Despite having fewer parameters, Qwen3-4B required more processing time than Gemma3-12B, suggesting that runtime was also influenced by model architecture and implementation.

GEC Model	Total (min)	Per Min (s)
Tiny-Aya-Fire (3.5B FT)	29.6	7.4
Gemma3-12B (Zero-Shot)	50.3	12.6
Qwen3-4B (FT)	61.6	15.4
Qwen3-14B (Zero-Shot)	94.9	23.7

Table 4: Inference time for the post-ASR correction stage.

4.6 Discussion

Taken together, the experimental results consistently show that fine-tuning small language models is more effective for Urdu post-ASR correction than relying on substantially larger models in zero-shot mode. Although the larger language models achieve stronger zero-shot performance, both fine-tuned 3.5–4B models outperform the 12–14B zero-shot baselines.

From a practical perspective, this framework simplifies real-world deployment. The underlying Whisper-Turbo recognizer remains unchanged, while improvements are obtained by updating only a lightweight text correction model. This avoids repeatedly adapting the ASR system itself and, as shown by the latency results above, introduces only a modest computational overhead while providing substantial gains in transcription quality.

Finally, the pseudo-gold data generation pipeline demonstrates that effective post-ASR correction models can be developed without manually annotating large correction corpora. By combining automatic transcription, quality filtering, and teacher-model correction, the proposed pipeline produces supervision of sufficient quality to fine-tune small language models. Although demonstrated here for Urdu, the same methodology is readily applicable to other low-resource languages facing similar data constraints.

5 Conclusion

This paper investigated Generative Error Correction (GEC) as a practical approach to improving Urdu automatic speech recognition without modifying the underlying ASR system. To address the

scarcity of manually corrected Urdu transcripts, we constructed a pseudo-gold corpus by correcting Whisper-Turbo transcriptions with Gemini and used it to fine-tune two small language models for post-ASR error correction, achieving substantial and statistically significant WER reductions over both the uncorrected baseline and larger zero-shot models.

Results suggest that targeted fine-tuning of small language models offers a practical, resource-efficient path to post-ASR correction in low-resource settings, without requiring costly ASR retraining or reliance on frontier-scale models. We release the pseudo-gold corpus and fine-tuned models to support future work on Urdu ASR and post-ASR correction.

Limitations

Although this work demonstrates clear improvements for Urdu post-ASR correction, a few practical limitations remain. First, text-based generative error correction operates only on the text output from ASR decoding. While language models effectively fix substitution errors, word segmentation, and phonetic misspellings, they cannot recover speech content lost due to severe acoustic deletions or truncation during initial ASR decoding. Resolving those structural deletion errors requires fine-tuning the underlying Whisper acoustic model itself.

Second, our evaluation relies on teacher-generated rather than independently human-transcribed references: both training labels and test-set targets are derived from Gemini corrections. We manually checked a 100-pair sample against the source audio and found Gemini’s corrections closely matched human judgment, but a larger and more systematic human evaluation would give a stronger guarantee of pseudo-gold quality. Relatedly, our train/test split is performed at the segment rather than source-video level, so segments from the same recording can appear in both partitions; a fully video-disjoint split, which we plan to adopt in future work, would give a more conservative estimate of generalization to unseen speakers.

Third, while our audio dataset includes diverse broad-domain sources like talk shows, reporter diaries, and podcasts, it consists mainly of standard Urdu speech from clear audio streams. Future work also includes evaluating this framework

on heavily accented regional dialects and noisy conversational speech, comparing against smaller encoder-decoder correctors, profiling end-to-end latency and memory beyond the correction stage, and extending the approach to additional low-resource languages.

Ethics Statement & Data Availability

To facilitate reproducibility and further research, we make our fine-tuned model adapters (Tiny-Aya-Fire⁸ and Qwen3-4B⁹) and the pseudo-gold transcription dataset¹⁰ publicly available on Hugging Face under the CC BY-NC 4.0 license. The original audio recordings and identifiers of the source YouTube videos are not distributed due to copyright and licensing considerations.

Acknowledgments

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

References

- Syed Ahad, Burhanuddin Aliasghar Ezzi, Muhammad Arsalan Hussain, Sandesh Kumar, and Abdul Samad. 2026. [Corpora generation for Urdu grammatical error correction](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 43428–43444, San Diego, California, United States. Association for Computational Linguistics.
- Zehra Ahmed, Farah Inayat, Zuha Aqib, and Sajjad Haider. 2026. Enhancing Urdu ASR with Whisper v3: Fine-tuning on latest datasets and realistic multi-speaker evaluation with SLM post-processing. In *Proceedings of the International Conference on Language Technologies for Low-Resource Languages (LaTeLL 2026)*, Fes, Morocco.
- Samee Arif, Sualeha Farid, Aamina Jamal Khan, Mustafa Abbas, Agha Ali Raza, and Awais Athar. 2025. WER we stand: Benchmarking Urdu ASR models. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 5952–5961.
- Sreyan Ghosh, Mohammad Sadegh Rasooli, Michael Levit, Peidong Wang, Jian Xue, Dinesh Manocha, and Jinyu Li. 2025. [Failing forward: Improving generative error correction for ASR with synthetic data](#)

⁸<https://huggingface.co/sajjadiba/urdu-asr-error-correction-tiny-aya-fire>

⁹<https://huggingface.co/sajjadiba/urdu-asr-error-correction-qwen3-4b>

¹⁰<https://huggingface.co/datasets/sajjadiba/urdu-asr-error-correction-data>

- and retrieval augmentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2466–2482, Vienna, Austria. Association for Computational Linguistics.
- Yun Hao, Reihaneh Amooie, Wietse de Vries, Rik van Noord, and Martijn Wieling. 2026. Can large language models reliably correct errors in low-resource ASR? A contamination-aware case study on West Frisian. *arXiv preprint arXiv:2605.19711*.
- Yuka Ko, Sheng Li, Chao-Han Huck Yang, and Tatsuya Kawahara. 2026. [Evaluating ASR-LLM setups for Japanese speech recognition with multi-pass augmented generative error correction](#). In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.
- Rishabh Kumar, Sabyasachi Ghosh, and Ganesh Ramakrishnan. 2024. [Beyond common words: Enhancing ASR cross-lingual proper noun recognition using large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6821–6828, Miami, Florida, USA. Association for Computational Linguistics.
- Rishabh Kumar, Amrith Krishna, Ganesh Ramakrishnan, and Preethi Jyothi. 2026. [Post-ASR correction in Hindi: Comparing language models and large language models in low-resource scenarios](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 636–645, Rabat, Morocco. Association for Computational Linguistics.
- Sheng Li, Chen Chen, Chin Yuen Kwok, Chenhui Chu, Eng Siong Chng, and Hisashi Kawai. 2024. [Investigating ASR error correction with large language model and multilingual 1-best hypotheses](#). In *Proceedings of Interspeech 2024*.
- 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*.
- Monika and Vishal Passricha. 2026. [LLM-based generative error correction for Hindi and Punjabi ASR using in-context learning and task-activating prompting](#). *International Journal of Advances in Signal and Image Sciences*, pages 1730–1744.
- Vahid Noroozi, Somshubra Majumdar, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2024. [Stateful conformer with cache-based inference for streaming automatic speech recognition](#). In *ICASSP 2024 - IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 12041–12045.
- 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 (ICML)*, volume 202, pages 28492–28518. PMLR.
- Alejandro R. Salamanca, Diana Abagyan, Daniel D’souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajae, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan Gomez, Ivan Zhang, and 7 others. 2026. [Tiny aya: Bridging scale and multilingual depth](#). *Preprint*, arXiv:2603.11510.
- Najm Ul Sehar, Ayesha Khalid, Farah Adeeba, and Sar-mad 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.
- Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. [Qwen3-asr technical report](#). *arXiv preprint arXiv:2601.21337*.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Natsuo Yamashita, Masaaki Yamamoto, Hiroaki Kokubo, and Yohei Kawaguchi. 2025. LLM-based generative error correction for rare words with synthetic data and phonetic context. *arXiv preprint arXiv:2505.17410*.
- Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. 2023. [Generative speech recognition error correction with large language models and task-activating prompting](#). In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

A Prompts

A.1 Teacher Correction Prompt (Gemini via OpenRouter)

Used during pseudo-gold dataset construction to generate corrected reference transcripts from Whisper-Turbo output.

PROMPT = """Correct the spelling/recognition errors in the given Urdu ASR transcript.

Output ONLY the corrected Urdu text. Do NOT add any explanations, comments, bullet points, markdown, or extra words.

Input: {input_text}
Corrected: """

A.2 Correction Model Prompt (Zero-Shot and Fine-Tuned Inference)

Used identically across zero-shot baselines (Tiny-Aya-Fire, Qwen3-4B-Instruct, Gemma3-12B, Qwen3-14B) and fine-tuned models (Tiny-Aya-Fire, Qwen3-4B-Instruct), both for training data formatting and at inference time. Finalized after iterative refinement to constrain output from small language models that otherwise deviated from simpler instructions.

```
ZERO_SHOT_PROMPT = """Task: Cor-
rect Urdu ASR errors.
Rules:
1. Fix ONLY spelling and character errors
(e.g., ~ ...).
2. Keep correct words EXACTLY as is —
DO NOT rewrite or paraphrase.
3. DO NOT add any notes, explanations, or
markers.
4. STOP immediately after the corrected
sentence.

Input: {input_text}
Corrected:"""
```