Corpora Generation for Urdu Grammatical Error Correction

Syed Ahad[†ζλ], **Burhanuddin Ezzi**[†ζλ], **Arsalan Hussain**[†ζλ], **Sandesh Kumar**[ζΩ], **Abdul Samad**[ζΩ]
Habib University[ζ], Karachi, Pakistan
{sa07753, be07724, mh07607}@st.habib.edu.pk[λ],
{sandesh.kumar, abdul.samad}@sse.habib.edu.pk[Ω]

Abstract

Grammatical Error Correction (GEC) for Urdu remains an under-researched area due to the lack of annotated datasets. This paper addresses the challenge of generating a robust corpus for fine-tuning deep learning models aimed at Urdu GEC. We propose a method for synthesizing a large dataset by collecting errors from the Urdu WikiEdits history, learning from them, and inserting similar errors in grammatically correct sentences to generate incorrect sentences with grammatical errors, hence creating a pair of grammatically correct and incorrect sentences. We introduce UrduGEC-Synthetic, a synthetically generated dataset produced through this pipeline[1]. Furthermore, we introduce UrduGEC-Gold, a Gold Dataset by extracting errors from exam copies of students[2]. Finally, we also fine-tuned various models on UrduGEC-Synthetic and evaluated them against UrduGEC-Gold to show the quality of synthetic data generation.

1 Introduction

The field of Grammatical Error Correction (GEC) has witnessed a paradigm shift in recent years, moving from early rule-based and statistical classifiers to deep learning approaches. Specifically, treating GEC as a sequence-to-sequence (Seq2Seq) translation task has become the standard in the field, largely following the pioneering work of Yuan and Briscoe (2016), who demonstrated the efficacy of Neural Machine Translation (NMT) architectures for correcting grammatical errors. However, the success of these supervised deep learning models is heavily contingent on the availability of massive, high-quality parallel corpora of grammatically incorrect and correct sentence pairs. While such resources exist for high-resource languages like English, they are virtually nonexistent for low-resource languages, creating a significant bottleneck for development.

Urdu is a language with a rich heritage and a script influenced by Persian, Arabic, Hindi, Sanskrit, and Turkish (Butt and Ahmed, 2011; Mangrio, 2016). It is widely spoken in South Asian countries, including Pakistan and India, by a large population. Despite its complexity and the variety of grammatical components, unfortunately, there has been little work done in the field of Urdu GEC. An automated GEC system for Urdu would greatly enhance content creation, digital communication tools, and educational platforms. To build a state-of-the-art model that generalizes well to a wide range of human grammatical errors and outputs correct sentences for grammatically incorrect inputs, we need a large corpus of training data.

To address this data scarcity, this paper proposes a robust method for synthetic data generation. We propose a pipeline that mines naturally occurring error patterns from Urdu Wikipedia revision histories and re-inflicts them onto clean text using basic morpho-syntactic analysis and widely available NLP tools, while avoiding manual annotation, proprietary resources, and language-specific handcrafted rules. Specifically, we ask: Can synthetic data generation for grammatical error correction of low resource Indic languages, operating under realistic low-resource constraints and assuming similar resource availability, achieve competitive or improved downstream performance compared to existing synthetic data generation paradigms while avoiding language-specific resources?

Finally, given the substantial morphosyntactic similarities Urdu shares with other Indo-Aryan languages such as Hindi, Punjabi, and Sindhi (Butt and Ahmed, 2011; Mangrio et al., 2021), we hypothesize that our data generation pipeline is transferable to this broader linguistic family. Provided that prerequisite resources such as reliable POS taggers and dependency treebanks are available, this methodology can be effectively adapted to develop robust GEC systems for other low-resource Indic languages.

[†] Equal contribution.

[0] Code for Synthetic Data Generation Pipeline: https://github.com/burhanuddin6/gec_pipeline.

[1] Synthetic dataset: https://huggingface.co/datasets/Kagura-Ahad-123/UrduGEC-Synthetic

[2] Gold dataset: https://huggingface.co/datasets/Kagura-Ahad-123/UrduGEC-Gold

Findings of the Association for Computational Linguistics: ACL 2026, pages 43428–43444

2 Related Work

Research in GEC has largely focused on methods to overcome the scarcity of annotated training data through synthetic error generation (Bryant et al., 2023). This section outlines the evolution of these techniques from random noise injection to linguistically grounded probabilistic modeling.

2.1 The Shift to Neural GEC and Data Scarcity

As established by Yuan and Briscoe (2016), modeling GEC as a monolingual machine translation task allows for powerful generalization, provided sufficient parallel data exists. Recognizing this acute data scarcity, Junczys-Dowmunt et al. (2018) successfully bridged the performance gap between neural and statistical GEC models by treating the task as a low-resource machine translation problem, utilizing techniques like transfer learning and source-side corruption. In the absence of large annotated corpora (such as NUCLE or Lang-8 used for English), researchers have increasingly turned to synthetic data generation. Early approaches, such as GenERRate by Foster and Andersen (2009), introduced probabilistic error-infliction tools to generate errors like POS insertions, deletions, and substitutions. While useful for small-to-medium corpora, purely random or simple rule-based noise often fails to capture the complexity of human errors, leading to models that do not generalize well to real-world inputs.

2.2 Linguistically Grounded Error Generation

To improve the quality of synthetic data, recent work has emphasized the importance of linguistic context over random noise. Sidorov et al. (2013) argued that traditional lexical n-grams are insufficient for modeling syntactic relations and demonstrated that Syntactic N-grams (utilizing POS tags) significantly improve error detection and correction. This theoretical foundation supports the move toward "kernel-based" or pattern-based error generation.

Building on this, Ma et al. (2022) utilized a Linguistic Rules-Based Generation (CLG) approach. They explicitly defined grammatical rules (e.g., "Structural Confusion") to generate synthetic errors that mimic the structural complexity of real mistakes. Their findings highlight that synthetically generated errors must be linguistically plausible to be effective for training NMT systems. This contrasts with earlier methods by Foster and Andersen (2009), which relied on hard-coded rules, by demonstrating that preserving the "grammatical severity" of an error is crucial for downstream model performance.

2.3 Mining Error Distributions from Natural Data

Rather than manually crafting linguistic rules, modern approaches attempt to learn error distributions directly from noisy data. Felice and Yuan (2014) were among the first to move away from inserting random errors, instead calculating conditional probabilities of errors based on linguistic context (e.g., $P(\text{error}|\text{POS tag})$) derived from learner corpora. Similarly, Kasewa et al. (2018) argued against random noise injection, proposing a method to learn the probability distribution of errors from a seed corpus and project that distribution onto clean text. Extending this paradigm to the neural era, Choe et al. (2019) employed ERRANT (Bryant et al., 2017) to automatically extract and categorize token-level edits from English learner essay corpora, building a noise dictionary of observed error types weighted by their empirical frequency. These categorized edits were then used to corrupt clean text for denoising pre-training, demonstrating that a high-quality annotated learner corpus, combined with a robust error-categorization tool, can yield a powerful and realistic noising function. This "mining and grafting" strategy ensures that the synthetic data reflects the natural distribution of errors made by humans, a technique we adopt in our pipeline by downsampling synthetic data to match natural error frequencies.

2.4 GEC for Low-Resource and Indic Languages

In the context of utilizing web-scale resources, Lichtarge et al. (2019) proposed generating corpora using Wikipedia edit histories (WikiEdits). They employed a dual strategy of probabilistic error introduction and Round-Trip Translation (RTT) through a high-resource bridge language. Their work demonstrated that Wikipedia revision histories contain valuable "natural" errors that benefit models significantly when fine-tuned.

Specific to Indic languages, Sonawane et al. (2020) adapted these strategies for Hindi, a language linguistically similar to Urdu. They generated a parallel corpus of synthetic errors by focusing on inflectional errors via a rule-based process and validated their models against scraped Wikipedia edit histories. Our work extends these methodologies to Urdu with two key distinctions. First, unlike Felice and Yuan (2014), who calculate probabilities from labeled learner corpora, we mine error patterns directly from noisy Wikipedia revision histories, removing the dependency on expensive human annotation. Second, distinct from Sidorov et al. (2013)

who use syntactic features for rule-based correction, we utilize local morpho-syntactic kernels to synthesize training data, enabling the neural models to learn complex inflectional nuances.

3 Dataset Generation

3.1 Methodology for Synthetically Generated Dataset

Our approach for creating a synthetically generated dataset is a pipeline that uses the established paradigm of "error annotation and infliction," a principled approach for creating synthetic data in low-resource settings (Alrehili and Alhothali, 2025). In the first phase of our pipeline (Annotation Phase), given a set of correct and grammatically incorrect sentence pairs, learns the different types of errors from those pairs, and in its second phase (Infliction Phase), given clean Urdu sentences, outputs corrupted sentences resulting in GEC pairs, where the corruption is the same as the one learned from the set of correct and grammatically incorrect sentence pairs.

3.1.1 Extraction of Errors from Wikipedia Revision Histories

To capture natural grammatical errors for the Annotation phase, we mined Urdu Wikipedia revision histories, adopting the strategy used by Sonawane et al. (2020) for Hindi. Given that Wikipedia content in low-resource languages often originates from translation followed by human post-editing, the edit history provides a rich source of grammatical corrections. We extracted edit sequences, treating the final revision as the grammatically correct reference and the penultimate revision as the incorrect source. Following extraction, we applied a standard cleaning pipeline to filter noise, the details of which are provided in Appendix A, which is also based on (Sonawane et al., 2020), ensuring that the errors collected are grammatical only. This process yielded the **Wiki-Edits Dataset**, comprising approximately 240,000 parallel sentence pairs.

3.1.2 Error Annotation

The goal of the annotation phase is to build a structured, context-rich database of observed grammatical errors from the **Cleaned Wiki-Edits Dataset** using each of its incorrect-correct sentence pairs.

First, we performed context-aware analysis using **Stanza** (Qi et al., 2020) to extract POS tags, lemmas (root words), and morphological features for every word in each sentence pair in the **Cleaned Wiki-Edits Dataset**. Stanza was selected for its high accuracy on Urdu (Appendix B). To ensure

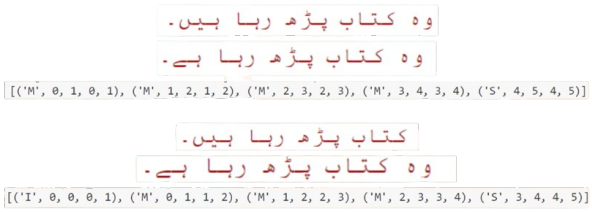

Figure 1: Two example outputs of the alignment algorithm on the same grammatically incorrect input sentence (the top one in both examples), which has five words in it; "وہ", "کتاب", "پڑھ", "رہا", and "ہیں". The output of the first example (transition of type S) shows that the sentence has 5 words in total, 4 matching words "M", and 1 substituted word "S". The output of the second example (transition of type S and I) shows that there has been an insertion "I" at the first index and a substitution at the last index.

that our pipeline focuses on morphological shifts rather than spelling corrections, we established a closed vocabulary for Out-of-Vocabulary (OOV) checks derived from the UD 2.12 treebank (Bhat et al., 2017; Palmer et al., 2009). This ensures tokenization consistency, as Stanza and our vocabulary share the same underlying schema from the treebank.

Next, we identified the precise edits transforming incorrect sentences into correct ones using a custom alignment algorithm inspired by **ERRANT** (Bryant et al., 2017), which classifies edits as 'Insertion' (I), 'Deletion' (D), or 'Substitution' (S); Figure 1 shows an example. Unlike standard Levenshtein distance, our algorithm employs a linguistically motivated cost function to better model grammatical similarity (exact formulation in Appendix C).

Kernel-Based Error Representation. To capture the grammatical environment of an error, we define a local context window, or "kernel," of size k. This consists of the token at the edit index (w_i) and its surrounding neighbors. For example, if $k = 3$, the kernel includes the immediate left and right neighbors (w_{i-1}, w_{i+1}). We treat k as a tunable hyperparameter to balance context-awareness against data sparsity. Rather than storing raw lexical forms, we abstract tokens into their **Morpho-Syntactic States** using Stanza. We observed that different error types require different levels of contextual granularity. Therefore, we defined distinct matching criteria for each operation:

- **Substitution Patterns:** We model this as a *Morphological Transition* of the target word constrained by a *Syntactic Frame*. The frame is defined strictly by the UPOS tags of the con-

text window. The transition is defined by the change in morphological features of the middle word (e.g., Case: $\mathrm{Nom} \rightarrow \mathrm{Case}$:cc). We do not constrain the morphological features of the surrounding tokens in the window for substitutions, allowing the pattern to generalize across different neighbor inflections.

- **Insertion and Deletion Patterns:** Since these operations involve the presence or absence of a word, we enforce a stricter context validation: the kernel is defined by both the UPOS tags *and* the exact morphological features of the surrounding words in the context window. For example, in the case of $k = 3$, deleting a postposition requires checking that the preceding noun is in the specific morphological state (e.g., Oblique) that licensed that postposition.

Annotation Storage. We aggregate these patterns into a database of observed errors, where each entry stores the matching criteria (Context UPOS and/or Features) and the transformation rules required to reproduce the error. For **Substitutions**, we record the specific mismatch in morphological features between the incorrect and correct words (e.g., a shift in Case). For **Deletions**, we store the exact surface word to be removed. For **Insertions**, we record the strict morphological context of the neighbors that indicates a missing word. Detailed schema definitions for these annotation objects are listed in Appendix D.

3.1.3 Error Infliction

In the final phase, we utilize the learned error patterns to corrupt the **Makhzan Corpus**[3], a large collection of cleaned, grammatically correct corpus consisting of around 270,000 Urdu sentences.

The infliction process operates by sliding a window across every sentence in the clean text, analyzing every possible local context window of size k with Stanza to extract UPOS tags and morphological features. A context window in the clean sentence is identified as a candidate for corruption if its local morpho-syntactic context exactly matches the *target* (correct) side of an error pattern stored in our Annotations Database.
To ensure high-quality generation, we apply specific linguistic constraints for each error type:

Substitution Infliction (The Lemma Constraint). Blindly substituting words based solely on POS tags would likely result in semantic drift. To prevent this, we generate substitution errors only when two conditions are met:

1. The word in the clean corpus must have morphological features matching the $\mathrm{correct\_feats}$ of a retrieved error pattern.
2. There must exist a derived form from the **lemma** of the clean word matching the $\mathrm{incorrect\_feats}$ of the pattern (verified via our **Lemma Dictionary**, see Appendix E.1).

These conditions ensure that the generated error is a valid morphological variant of the original word, preserving the sentence's semantic core.

Insertion and Deletion Infliction. For these operations, we apply the logic of "reversing the correction" with strict context validation (described in Appendix E.2):

- **Deletions:** To simulate a deletion error, we identify a context in the clean sentence that matches a learned *Insertion* pattern. If the clean words match the full morphological profile of an inserted word and its neighbors, the middle word is removed to generate the erroneous sentence of the pair.
- **Insertions:** To simulate an insertion error, we look for a context matching a learned *Deletion* pattern. If a sequence of clean words matches the context defined by a Deletion pattern (where the 'gap' represents the token to be inserted), we inject the missing word as stored in the annotation database to generate the erroneous sentence of the pair.

3.2 Creation of the UrduGEC-Gold Dataset

While synthetic data is essential for training in low-resource settings, evaluating models solely on synthetic data often leads to inflated performance metrics that fail to reflect real-world generalization. To establish a rigorous, unbiased benchmark for Urdu GEC, we introduce **UrduGEC-Gold**, a high-quality Gold test dataset of naturally occurring errors.

Data Sources and Curation. We collaborated with local educational institutions to procure work from Urdu exam papers of students (Grade 8 and above). The selected schools were chosen so that the student demographic there comprises both L1 and L2 learners to add diversity. Additionally, we extracted exercises from Urdu grammar workbooks designed for error correction tasks.

[3] https://github.com/zeerakahmed/makhzan/

Scope and Filtering. We used help from expert linguists to manually filter out purely orthographic errors (spelling mistakes) or stylistic fluency edits that did not violate grammatical rules. This resulted in a dataset of 1,613 high-quality parallel pairs.

Error Distribution Statistics. To provide transparency regarding the types of errors covered, we provide statistics of UrduGEC-Gold using the standardized ERRANT-type metrics (Bryant et al., 2017) to illustrate the distribution of error types. We group the errant type metrics into 5 categories detailed in Appendix H. Table 1 details the breakdown of error types.

Error Category	Count	%	Example (Incorrect → Correct)
Verb & Aux	1,131	33.2	کرتی → کرتا (Gender/Number)
Noun & Pron	733	21.5	لڑکے → لڑکا (Oblique Case)
Adpos	560	16.4	کی → کا (Genitive Agreement)
Mod & Misc	618	18.2	اچھے → اچھا (Adj Agreement)
Ortho	362	10.6	مسئلہ → مسئلہ (Spelling)
Total Edits	**3,404**	**100.0**	**Avg. 2.1 Errors / Sentence**

Table 1: Distribution of error types in UrduGEC-Gold (1,613 sentences), grouped by macro-grammatical categories.

UrduGEC-Gold is, to our knowledge, the first expert-annotated GEC benchmark for Urdu derived from diverse educational sources.

Evaluation Protocol. This dataset is treated exclusively as a **Held-Out Test Set**. It is never seen by the model during any training experiment. By fine-tuning on synthetic data and evaluating on this completely different distribution of human errors, we ensure that our reported metrics reflect genuine grammatical generalization rather than pattern memorization.

4 Optimization of Synthetic Data Generation

The methodology described in Section 3.1 allows for flexible generation of errors based on three configurable parameters:

1. **Context Window Size (k):** Defines the granularity of the morpho-syntactic environment required to trigger an error injection.
2. **Sampling Strategy:** Determines the probability distribution used to select which error pattern to inflict when multiple matching patterns exist.
3. **Error Density:** Controls the number of distinct errors injected into a single sentence.

To determine the optimal configuration for generating the final synthetic corpus, we conducted a series of ablation studies.

4.1 Optimization Protocol

For these optimization experiments, we utilized the **mT0-large** (1.2B parameters) (Muennighoff et al., 2022) model as a proxy. We fine-tuned the model on different synthetic variations and evaluated performance on UrduGEC-Gold. All ablation runs used identical hyperparameters (detailed in Appendix F) to ensure that performance differences were solely attributable to data characteristics.

Metrics: To quantify performance, we utilize established metrics (Yuan and Briscoe, 2016) for Grammatical Error Correction:

- **MaxMatch (M^2):** We use the M^2 scorer (Dahlmeier and Ng, 2012) (with Stanza for tokenization) to report **Precision**, **Recall**, and **$F_{0.5}$**.
- **GLEU:** The Generalized Language Evaluation Understanding metric (Napoles et al., 2016), utilized to measure fluency and overlap with reference sentences.

Significance Testing: To ensure the robustness of our findings, we evaluate the statistical significance of performance differences between models using paired bootstrap resampling (Koehn, 2004) with $N = 1000$ iterations. Throughout this paper, we denote statistically significant improvements over the respective baseline with † for $p < 0.05$ and ‡ for $p < 0.01$.

4.2 Ablation 1: Context Window Size (k)

The kernel size k dictates the granularity of the error context. While a larger k theoretically captures longer-range dependencies, it increases data sparsity, as finding larger exact n-gram matches in a clean corpus becomes exponentially harder.

We experimented with $k = 3$ (± 1 neighbor), $k = 5$ (± 2), and $k = 7$ (± 3). For this experiment, we held the **Sampling Strategy (Natural)** and **Error Density (Single Edit)** constant.

Kernel	Pairs	P	R	$F_{0.5}$	GLEU
$k = 3$ (± 1)	1,270,499	**36.74**	**15.52**	**28.85**‡	**68.6**
$k = 5$ (± 2)	156,085	28.26	12.04	22.26	67.05
$k = 7$ (± 3)	18,689	27.84	08.44	19.07	67.3

Table 2: Impact of Kernel Size on GEC performance. Dataset size represents the number of sentences where a matching context was found in the clean corpus. ‡$p < 0.01$ against $k = 5$ and $k = 7$.

As shown in Table 2, increasing the kernel size

led to a linear degradation in performance. This is directly correlated with the massive reduction in yieldable data; $k = 7$ produced only $\sim$1.5% of the data volume compared to $k = 3$. Consequently, we selected $k = 3$ as the optimal trade-off between local context awareness and data availability.

4.3 Ablation 2: Error Sampling Strategy

Errors in the Wiki-Edits dataset follow a "Natural" long-tail distribution, where simple errors (e.g., adposition drops) are vastly more frequent than complex morphological shifts. We compared **Natural Sampling** against **Temperature Sampling** ($\tau = 0.5$), which flattens the distribution to give rare error patterns higher representation.

For this experiment, we held **Kernel ($k = 3$)** and **Error Density (Single Edit)** constant. Crucially, to ensure a fair comparison, the dataset size for Temperature Sampling was deliberately constrained to match the Natural Sampling size ($\approx$ 1.27M sentences) using similar implementation to logic detailed in Appendix G.

Strategy	P	R	$F_{0.5}$	GLEU
Natural Sampling	36.74	**15.52**	28.85	68.6
Temperature Sampling	**40.50**	15.00	**30.22**†	**69.08**

Table 3: Impact of Sampling Strategy. Both datasets contained $\approx$ 1.27M sentences. $^{\dagger}p < 0.05$ against Natural Sampling.

Temperature sampling yielded a statistically significant improvement in Precision and $F_{0.5}$ (Table 3). By upsampling rarer morphological error patterns, the model learned to generalize better to diverse error types, rather than overfitting to high-frequency surface edits.

4.4 Ablation 3: Error Density (Single vs. Multi-Edit)

Real-world texts often contain multiple errors per sentence. To support **Multi-Edit** generation, we engineered our infliction pipeline to apply errors iteratively from **right-to-left** to preserve index integrity. We compared a dataset with exactly one error per sentence against one where the number of errors follows a normal distribution using mean and standard deviation derived from Gold statistics (inspired by Grundkiewicz et al. (2019)).

For this experiment, we held **Kernel ($k = 3$)** and **Sampling Strategy (Temperature)** constant.

Even though UrduGEC-Gold contains a significant proportion of multi-edit sentences (789 multi-edit vs. 824 single-edit pairs), the Multi-Edit model performed worse (Table 4). We hypothesize that this is due to **infliction intensity**: applying a normal distribution of errors to the relatively short sentences typical of our corpus often obscures the semantic context required for reconstruction. Facing such high noise levels during training, the model struggles to learn precise correction mappings, leading to a notable drop in Precision (from 40.50 to 30.61). Consequently, we utilized the **Single Edit** configuration for the final corpus.

Configuration	P	R	$F_{0.5}$	GLEU
Single Edit	**40.50**	**15.00**	**30.22**‡	**69.08**
Multi-Edit (Normal Dist.)	30.61	09.76	21.45	67.6

Table 4: Impact of Error Density on GEC performance. $^{\ddagger}p < 0.01$ against Multi-Edit.

4.5 Resulting synthetic GEC Corpora

Based on the optimization study, we generated our final synthetic training corpus using **Kernel** $k = 3$, **Temperature Sampling**, and **Single Edit** injection of length 1.27 million sentence pairs. We name it UrduGEC-Synthetic.

To ensure our generation methodology successfully captured the linguistic properties of natural errors, we analyzed the distribution of error types within this final synthetic corpus. As detailed in Table 5, the resulting error distribution roughly mirrors the morpho-syntactic focus observed in UrduGEC-Gold (Table 1), maintaining a heavy concentration on verbal complexes and nominal inflections.

Error Category	Count	%
Noun & Pron	534,990	42.11
Verb & Aux	326,889	25.73
Adpos	241,476	19.01
Mod & Misc	141,136	11.11
Ortho	26,008	2.05
Total	**1,270,499**	**100.00**

Table 5: Distribution of error types in the final 1.27M UrduGEC-Synthetic Corpus. The distribution roughly mirrors UrduGEC-Gold's morpho-syntactic focus.

5 Experiments and Results

5.1 Baselines and Comparisons

To validate the efficacy of our proposed **Wiki-Edit Inflection** pipeline, we compare it against three distinct baselines using the same heuristics defined in Section 4.1 (Metrics and Significance Testing), which are regarded as state-of-the-art in synthetic error generation:

Zero-Shot Baseline: The raw pre-trained model

without any GEC-specific fine-tuning, serving as a control to measure inherent rewriting ability.

Random Noise Injection: Treating GEC as a denoising task has proven effective in low-resource NMT (Junczys-Dowmunt et al., 2018). We adapted Grundkiewicz et al. (2019)'s strategy, generating 1.27M sentence pairs via probabilistic edit operations (insertion, deletion, substitution, swap) on the Makhzan corpus, using Levenshtein distance for substitutions in lieu of Urdu spell-checking resources (Appendix G). This tests whether our linguistically motivated kernels provide value over stochastic noise.

Neural AEG (Back-Translation): An effective data augmentation technique in low-resource GEC (Bryant et al., 2023), following Koyama et al. (2021), Rei et al. (2017), and Htut and Tetreault (2019). We fine-tuned **mT0-large** in the *Correct* $\rightarrow$ *Incorrect* direction on raw Wiki-Edits data, then generated five predictions per sentence (Xie et al., 2018) to match our dataset size. We exclude generative models (Luhtaru et al., 2024) for fair comparison, and hypothesize the low scores result from unfiltered non-grammatical edits in raw Wiki-Edits data.

Raw Wiki-Edits: Fine-tuning directly on the $\sim$240K Wiki-Edits sentence pairs (§3.1.1), testing whether synthetic amplification provides gains over the natural source data.

Approach	P	R	$F_{0.5}$	GLEU
Zero-Shot	21.52	**25.08**	22.15	45.28
Random Noise	27.70	09.24	19.79	67.62
Neural AEG	10.51	03.04	07.05	62.58
Raw Wiki-Edits	20.74	05.40	13.22	66.02
Our Approach	**40.50**	15.00	**30.22**‡	**69.08**

Table 6: Our approach compared with baselines. $^{\ddagger}p < 0.01$ against all baselines.

An analysis of the results in (Table 6) highlights several key advantages of our proposed methodology:

- Our method achieves significantly higher Precision (40.50) vs. Raw Wiki-Edits (20.74) and Neural AEG (10.51), validating that kernel-based mining successfully filters non-grammatical noise that confuses the baseline approaches.
- Random Noise Injection yields high GLEU but very low Recall (9.24), suggesting it teaches the model to smooth text without learning specific Urdu morphological rules (e.g., case agreement), which our targeted inflection pipeline captures.
- The low performance of Neural AEG confirms that back-translation methods are highly sensitive to seed corpus quality; without our prefiltering, Neural AEG merely amplifies the noise present in raw Wikipedia revisions.
- The Recall drop stems from our pipeline's conservative filtering of noisy Wikipedia edits, which prioritizes morpho-syntactic fixes over broader coverage, producing a high-Precision but lower-Recall corrector. Given that $F_{0.5}$ favors Precision, this trade-off does not substantially harm overall performance.

5.2 Models and Implementation

We evaluated our finalized UrduGEC-Synthetic dataset on three competitive multilingual architectures to demonstrate model-agnostic gains:

- **mT0-large** (1.2B parameters): An instruction-tuned variant of mT5. (Muennighoff et al., 2022)
- **ByT5-base** (582M parameters): A token-free model operating on raw bytes, theoretically advantageous for morphological errors. (Xue et al., 2022)
- **NLLB-3.3B**: A massive multilingual translation model. (Team et al., 2022)

Hyperparameters and evaluation metrics are detailed in Appendix F and Section 4.1, respectively.

5.3 Main Quantitative Results

To measure the impact of our UrduGEC-Synthetic corpus, we evaluated each model on UrduGEC-Gold in two settings: a zero-shot baseline (before fine-tuning) and after fine-tuning on our 1.27 million sentence pairs. The results are summarized in Table 7.

The results reveal a substantial and consistent improvement across all metrics for both models after being fine-tuned on our UrduGEC-Synthetic corpus. The low baseline scores, particularly for NLLB, confirm that even powerful multilingual models lack specialized GEC capacity for Urdu out-of-the-box. Fine-tuning provides a dramatic performance boost.

5.4 LLM-as-a-Judge Evaluation

To validate correction quality beyond surface-level metrics, we adopt an LLM-as-a-Judge framework

Model	Variant	Error Categories											
		Verb & Aux		Noun & Pron		Adpos		Mod & Misc		Ortho		Overall	
		$F_{0.5}$	GLEU	$F_{0.5}$	GLEU	$F_{0.5}$	GLEU	$F_{0.5}$	GLEU	$F_{0.5}$	GLEU	$F_{0.5}$	GLEU
MT0	large (baseline)	24.27	45.35	16.35	43.71	27.10	51.21	22.20	47.74	13.44	36.20	22.15	45.28
	large (finetuned)	32.74	67.62	18.56	66.35	49.86	78.03	8.36	73.89	23.20	70.63	**30.22**‡	69.08
ByT5	base (baseline)	7.53	15.92	3.69	14.70	7.13	17.87	5.50	18.29	3.99	14.50	6.19	15.95
	base (finetuned)	29.11	68.16	14.93	66.99	43.03	77.00	5.13	76.93	8.82	71.17	**25.87**‡	69.66
NLLB	3.3B (baseline)	7.12	23.29	3.88	23.07	11.05	23.33	11.38	22.72	2.07	30.19	7.00	23.51
	3.3B (finetuned)	29.86	66.04	16.84	62.83	37.68	74.48	14.29	71.73	13.01	66.43	**26.12**‡	66.66

Table 7: Category-wise $F_{0.5}$ and GLEU scores on UrduGEC-Gold. Finetuned models demonstrate substantial gains across all categories compared to zero-shot baselines. $^{\ddagger}p < 0.01$ for Overall $F_{0.5}$ compared to the respective model baseline.

Metric	Model	Variant	Verb & Aux	Noun & Pron	Adpos	Mod & Misc	Ortho	Overall
Grammaticality	mT0-large	baseline	**88.75**	**87.76**	**87.42**	**86.61**	**84.70**	**89.04**
		finetuned	80.48	80.63	80.87	79.23	74.63	83.10
	ByT5-base	baseline	25.27	25.70	26.20	26.45	25.37	25.85
		finetuned	**68.30**	**70.49**	**67.81**	**68.63**	**65.07**	**72.16**
	NLLB-3.3B	baseline	27.54	28.14	27.42	26.72	26.79	27.91
		finetuned	**83.08**	**82.07**	**82.00**	**81.75**	**75.07**	**84.89**
Fluency	mT0-large	baseline	**85.4**	**82.9**	**84.5**	**82.6**	**81.0**	**85.7**
		finetuned	80.7	80.3	80.9	78.7	76.3	83.4
	ByT5-base	baseline	22.3	22.7	23.6	23.1	22.4	22.8
		finetuned	**68.8**	**70.4**	**67.4**	**68.1**	**66.0**	**72.4**
	NLLB-3.3B	baseline	21.1	21.2	21.1	21.2	21.3	21.3
		finetuned	**84.0**	**83.1**	**82.3**	**82.0**	**77.4**	**85.8**
Meaning Pres.	mT0-large	baseline	77.5	76.4	77.2	73.1	72.4	76.5
		finetuned	**99.0**	**98.8**	**99.2**	**99.0**	**98.7**	**99.0**
	ByT5-base	baseline	39.7	41.0	40.6	41.3	38.7	41.2
		finetuned	**89.3**	**88.9**	**86.9**	**87.1**	**84.9**	**90.7**
	NLLB-3.3B	baseline	57.6	59.6	58.4	59.8	60.4	58.1
		finetuned	**98.1**	**98.0**	**98.5**	**98.2**	**97.9**	**98.1**
Edit Difficulty	mT0-large	baseline	**52.2**	**50.2**	**50.8**	**49.1**	**49.5**	**49.8**
		finetuned	41.0	44.4	44.0	42.8	34.8	44.5
	ByT5-base	baseline	20.1	20.1	20.1	20.1	20.1	20.1
		finetuned	**34.1**	**38.4**	**36.2**	**36.2**	**29.9**	**37.6**
	NLLB-3.3B	baseline	20.3	20.3	20.1	20.3	20.3	20.2
		finetuned	**42.9**	**43.5**	**43.6**	**44.4**	**34.2**	**44.0**
Edit Impact	mT0-large	baseline	**62.4**	**57.7**	**58.9**	**56.0**	**52.6**	**59.8**
		finetuned	56.9	51.4	55.3	49.3	41.4	56.6
	ByT5-base	baseline	20.4	20.1	20.2	20.1	20.1	20.2
		finetuned	**43.2**	**41.9**	**40.2**	**39.5**	**32.8**	**43.7**
	NLLB-3.3B	baseline	20.3	20.2	20.1	20.4	20.3	20.2
		finetuned	**62.4**	**55.7**	**58.3**	**56.1**	**43.9**	**60.7**

Table 8: LLM-as-a-Judge qualitative scores across macro error categories. The scores are evaluated on a 1-5 scale and multiplied by 20 to be reported as percentages (0-100) for clarity.

(Kobayashi et al., 2024), using Gemini 3 Flash as the judge model — a highly capable, reasoning model well-suited for structured evaluation tasks (Google DeepMind, 2025). We evaluate our models across five criteria: Grammaticality, Fluency, Meaning Preservation, Edit Difficulty, and Edit Impact (full prompts in Appendix I).

Table 8 presents the qualitative evaluation scores. Fine-tuned **ByT5** and **NLLB** models demonstrate dramatic improvements across almost all metrics compared to their zero-shot baselines. In contrast, **mT0** exhibits a more nuanced shift: as a heavily instruction-tuned model, fine-tuning drives it toward conservatism, achieving near-perfect Meaning Preservation (≈99%) but with slight decreases in the other metrics. This reflects a "copy heuristic" rather than a degradation in capability, and aligns precisely with our quantitative finding of high Precision but lower Recall.

6 Conclusion

This paper addressed the critical scarcity of resources for Urdu Grammatical Error Correction (GEC) by introducing a principled, two-phase pipeline for large-scale synthetic corpus generation. Our approach learns fine-grained, context-aware er-

ror patterns from Wikipedia revision histories and systematically inflicts them onto a clean monolingual corpus, resulting in two key contributions: a replicable generation methodology and the release of a 1.27 million-pair synthetic corpus (UrduGEC-Synthetic) and a ”Gold” test set (UrduGEC-Gold). The profound effectiveness of this approach is confirmed by our experiments, which show dramatic performance gains across state-of-the-art models like mT0 and NLLB after fine-tuning. These results validate our method and provide the essential contributions to spur the development of practical GEC tools for millions of Urdu speakers. Furthermore, given the substantial morphosyntactic similarities Urdu shares with other Indo-Aryan languages such as Hindi, Punjabi, and Sindhi (Butt and Ahmed, 2011; Mangrio et al., 2021), we hypothesize that our data generation pipeline is transferable to this broader linguistic family. Provided that prerequisite resources such as reliable POS taggers and dependency treebanks are available, this methodology can be effectively adapted to develop robust GEC systems for other low-resource Indic languages.

Limitations

While our work establishes a strong baseline, we acknowledge several limitations that offer avenues for future research:

- **Source Data Domain:** Although the WikiEdits corpus provides a rich source of naturalistic errors, its encyclopedic style may not fully capture the nuances and error patterns common in other domains, such as conversational text or creative writing. The error distribution is primarily reflective of Wikipedia contributors, which may differ from that of language learners.
- **Scope of Error Correction:** Our kernel-based approach, which relies on a three-word local context, is highly effective for single-token edits (substitutions, insertions, deletions) but may not adequately capture more complex, non-local grammatical errors, such as incorrect sentence structure or discourse-level phenomena.
- **Absence of a Language-Specific Error Categorization Tool:** A direct experimental comparison with approaches that rely on ERRANT-style error categorization – such as Choe et al. (2019) – is currently precluded by the lack of an ERRANT-equivalent for Urdu. Such methods critically depend on this tool to extract and categorize edits from a seed corpus into linguistically typed error classes. Developing a dedicated Urdu or multilingual alternative to ERRANT would make these comparisons feasible, and thus serving as rigorous apples-to-apples baselines.
- **Evaluation on Limited Gold Data:** Due to the scarcity of annotated resources for Urdu, our ”Gold” test set, UrduGEC-Gold, while high-quality, is of a limited size. A larger, more diverse evaluation benchmark would be needed to make more definitive claims about real-world performance across different demographics of Urdu speakers.
- **Computational Constraints:** Fine-tuning large models required significant computational resources, which constrained our ability to conduct more extensive hyperparameter searches or train for extended epochs. This may have prevented some models from reaching their full potential on our dataset.

Ethical Statement

We have considered the ethical implications of our work throughout the research process.

- **Data Sources and Privacy:** The primary source for our synthetic data generation, Wikipedia, is a public resource with content licensed under Creative Commons (CC BY-SA 3.0). The text is encyclopedic and generally non-personal. While we did not implement an explicit PII removal step, a manual review of a sample of the data did not reveal any instances of sensitive personal information. The usernames associated with edits are pseudonymous and were not used in any part of our analysis. The UrduGEC-Gold dataset was derived from anonymized student exam papers, collected with institutional consent for research purposes. All personally identifiable information was removed during the manual extraction process.
- **Potential for Bias:** Our UrduGEC-Synthetic dataset inherits the biases present in its source corpora. The ”Wiki-Edits Dataset” reflects the language and error patterns of its contributors, while UrduGEC-Gold reflects those of students in a specific educational context. The models trained on this data will consequently learn these biases. We believe that releasing both datasets allows future researchers to

study and mitigate these biases, but users of models trained on this data should be aware of its origins.

- **Intended Use and Impact:** The intended use of our work is to advance NLP research and tool development for Urdu, a low-resource language. An automated GEC system has a significant positive impact by enhancing content creation, improving digital communication tools, and providing valuable feedback in educational platforms. We are not aware of any direct potential for malicious use of this technology. Both the code and the datasets are publicly available to ensure transparency, reproducibility, and to foster further research for the benefit of the Urdu-speaking community.

Acknowledgements

We would like to express our sincere gratitude to Ms. Sarwat Mifzal for her crucial assistance in connecting us with local educational institutions to procure the student exam copies for our UrduGEC-Gold dataset, as well as for her valuable insights during the ideation of this project. We are equally grateful to Munawwar Anwar, Huzaifa Riaz, and Ahmed Ali from Habib University for their thoughtful discussions and contributions towards analyzing the core ideas presented in this work. Finally, we extend our heartfelt thanks to Huzaifa Riaz, Ahmed Ali, and Roshaan Khan from Habib University for their diligent efforts in digitizing the gold data from the physical exam copies. Their support was instrumental in bringing this research to fruition.

References

Ahlam Alrehili and Areej Alhothali. 2025. Towards the development of balanced synthetic data for correcting grammatical errors in arabic: An approach based on error tagging model and synthetic data generating model. *Preprint*, arXiv:2502.05312.

Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, and Fei Xia. 2017. The hindi/urdu treebank project. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 659–697. Springer Netherlands, Dordrecht.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3):643–701.

Miriam Butt and Tafseer Ahmed. 2011. The redevelopment of indo-aryan case systems from a lexical semantic perspective. *Morphology*, 21(3):545–572.

Yo Joong Choe, Jiyeon Ham, Kyubyong Park, and Yeoil Yoon. 2019. A neural grammatical error correction system built on better pre-training and sequential transfer learning. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–227, Florence, Italy. Association for Computational Linguistics.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Mariano Felice and Zheng Yuan. 2014. Generating artificial errors for grammatical error correction. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–126, Gothenburg, Sweden. Association for Computational Linguistics.

Jennifer Foster and Oistein Andersen. 2009. GenERRate: Generating errors for use in grammatical error detection. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 82–90, Boulder, Colorado. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 3 Flash Model Card. Technical report, Google DeepMind.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263, Florence, Italy. Association for Computational Linguistics.

Phu Mon Htut and Joel Tetreault. 2019. The unbearable weight of generating artificial errors for grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 478–483, Florence, Italy. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*,

pages 595–606, New Orleans, Louisiana. Association for Computational Linguistics.

Sudhanshu Kasewa, Pontus Stenetorp, and Sebastian Riedel. 2018. Wronging a right: Generating better errors to improve grammatical error detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4977–4983, Brussels, Belgium. Association for Computational Linguistics.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. Large language models are state-of-the-art evaluator for grammatical error correction. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 68–77, Mexico City, Mexico. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Aomi Koyama, Kengo Hotate, Masahiro Kaneko, and Mamoru Komachi. 2021. Comparison of grammatical error correction using back-translation models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 126–135, Online. Association for Computational Linguistics.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 2019. Corpora generation for grammatical error correction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3291–3301, Minneapolis, Minnesota. Association for Computational Linguistics.

Agnes Luhtaru, Taido Purason, Martin Vainikko, Maksym Del, and Mark Fishel. 2024. To err is human, but llamas can learn it too. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12466–12481, Miami, Florida, USA. Association for Computational Linguistics.

Shirong Ma, Yinghui Li, Rongyi Sun, Qingyu Zhou, Shulin Huang, Ding Zhang, Li Yangning, Ruiyang Liu, Zhongli Li, Yunbo Cao, Haitao Zheng, and Ying Shen. 2022. Linguistic rules-based corpus generation for native Chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 576–589, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Riaz Ahmed Mangrio. 2016. *The Morphology of Loanwords in Urdu: The Persian, Arabic and English Strands*. Cambridge Scholars Publishing, Newcastle upon Tyne.

Riaz Ahmed Mangrio, Muhammad Javed Iqbal, Zafeer Hussain Kiani, and Rashida Imran. 2021. Morpho-phonological similarities in indo aryan languages: A descriptive account. *Kashmir Journal of Language Research*, 23(2).

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2016. Gleu without tuning. *Preprint*, arXiv:1605.02592.

Martha Palmer, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. Hindi syntax: Annotating dependency, lexical predicate-argument structure, and phrase structure. In *The 7th International Conference on Natural Language Processing*, pages 14–17.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Marek Rei, Mariano Felice, Zheng Yuan, and Ted Briscoe. 2017. Artificial error generation with machine translation and syntactic patterns. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 287–292, Copenhagen, Denmark. Association for Computational Linguistics.

Grigori Sidorov, Anubhav Gupta, Martin Tozer, Dolors Catala, Angels Catena, and Sandrine Fuentes. 2013. Rule-based system for automatic grammar correction using syntactic n-grams for English language learning (L2). In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 96–101, Sofia, Bulgaria. Association for Computational Linguistics.

Ankur Sonawane, Sujeet Kumar Vishwakarma, Bhavana Srivastava, and Anil Kumar Singh. 2020. Generating inflectional errors for grammatical error correction in Hindi. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 165–171, Suzhou, China. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Ng, and Dan Jurafsky. 2018. Noising and denoising natural language: Diverse backtranslation for grammar correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 619–628, New Orleans, Louisiana. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386, San Diego, California. Association for Computational Linguistics.

Appendices

A Raw Wiki-Edits Cleaning

To ensure the quality of the "Wiki-Edits Dataset," the raw data extracted from Wikipedia revision histories were subjected to the following cleaning pipeline:

1. **Character Filtering:** Removal of extra whitespaces and non-Urdu characters (e.g., emojis or unsupported symbols).
2. **Noise Filtering:** Removal of sequences containing a high percentage of numeric or non-alphabetical characters, as these usually represent metadata or formatting code rather than natural language.
3. **Deduplication:** Removal of extremely similar sentence pairs to prevent data leakage and redundancy.
4. **Unicode Normalization:** Translation of non-Urdu Unicode characters to the standard Urdu range (e.g., converting Arabic-specific Unicode points to their Urdu equivalents) and translation of English numerals to Urdu numerals using the stand-alone normalization method provided by the UrduHack Toolkit[4].

B Linguistic Resources

Stanza's Urdu model, fine-tuned on the UD 2.12 treebank dataset (Bhat et al., 2017; Palmer et al., 2009), demonstrates high performance, with reported accuracy scores of 94.78% for lemmatization, 92.98% for UPOS tagging, and 84.06% for morphological features. This level of accuracy is crucial for the annotation phase, as it ensures that the grammatical context of each error is captured with high fidelity.

C Alignment Cost Function

Our alignment algorithm utilizes a weighted cost function to prioritize grammatical changes over semantic substitutions. The cost between an original token o and a corrected token c is calculated as:

$$\text{Cost}(\text{o}, \text{c}) = \text{Cost}_{\text{lemma}} + \text{Cost}_{\text{pos}} + \text{Cost}_{\text{char}}$$

Where:

- $\text{Cost}_{\text{lemma}}$ is 0 if the lemmas are identical and 0.499 otherwise. This penalizes substitutions that change the base meaning of the word.[5]
- Cost_{pos} is 0 if the Universal Part-of-Speech (UPOS) tags are the same, 0.25 if both are open-class POS tags (e.g., Noun to Verb), and 0.5 otherwise. This reflects the grammatical severity of the change.
- $\text{Cost}_{\text{char}}$ is the normalized Indel distance between the token accounting for minor spelling variations.

This cost function ensures that the alignment prefers linguistically plausible edits, such as changing an inflection, over substituting an unrelated word.

D Detailed Annotation Schema

This appendix details the data structures used to implement the kernel-based error representation described in Section 3.1.2. The schema is designed to be flexible for any kernel size k, though we illustrate using the default $k = 3$.

D.1 Kernel Definition and Placeholders

To handle boundary conditions, the "kernel" utilizes a specific placeholder tag, '%', to handle edge cases:

- **Standard Context:** For Substitution and Insertion errors, the kernel is a list of k UPOS tags (e.g., [UPOS_left, UPOS_middle, UPOS_right]). If the edit occurs at the start or end of a sentence, the missing neighbor is replaced with '%'.
- **Deletion Context:** Since a Deletion error represents a "gap" between two words, the middle element is always the placeholder. The structure is strictly [UPOS_left, '%', UPOS_right] (for $k = 3$).

D.2 JSON Annotation Objects

For every observed edit, a structured JSON object is created. The specific fields vary by error type to optimize for the distinct context requirements of each operation (e.g., storing neighbor features for Indels but not for Substitutions).

Substitution (S) An annotation is created only if both the incorrect and correct words pass the OOV check.

- type: 'S'
- kernel_upos: The list of three UPOS tags forming the syntactic frame.
- incorrect_feats: A UPOSFeats object containing the UPOS and morphological features

[4] https://github.com/urduhack/urduhack/blob/master/urduhack/normalization/character.py

[5] 0.499 is chosen to be just below the threshold of 0.5 to prioritize lemma matches over POS mismatches in edge cases

(e.g., Gender=Masc|Number=Sing) of the **source** word.
- correct_feats: A UPOSFeats object containing the features of the **target** word.
- occurrence: A counter for this specific morphological transition under this kernel.

Deletion (D) An annotation is created if the deleted word and its neighbors pass the OOV checks. Note that, unlike Substitution, we store the morphological features of the *neighbors* to ensure strict context matching.

- type: 'D'
- kernel_upos: The three-element kernel with '%' in the middle.
- kernel_feats: A list of UPOSFeats objects for the left and right context neighbors.
- deleted_words: The list containing the exact surface word that was deleted.
- occurrence: Frequency counter.

Insertion (I) An annotation is created if the inserted word and the surrounding context in the correct sentence pass the OOV check.

- type: 'I'
- kernel_upos: The list of three UPOS tags.
- kernel_feats: A list of UPOSFeats objects for all three words (Left, Inserted, Right) to capture the full morphological agreement context.
- inserted_word: The exact surface word that was inserted.
- occurrence: Frequency counter.

D.3 Storage Structure

To allow for efficient querying during the infliction phase, these objects are stored in a dictionary keyed by the syntactic frame:

- **Key:** The string representation of the kernel_upos list (e.g., "['NOUN', '%', 'VERB']").

- **Value:** A list of all unique annotation objects (as defined above) that share that syntactic frame.

Figure 2 visualizes how these objects are serialized in the final JSON output.


```
{
  "['PRON', 'NOUN', 'ADP']": [
    {
      "type": "S",
      "kernel_upos": ["PRON", "NOUN", "ADP"],
      "incorrect_feats": {
        "upos": "NOUN",
        "feats": "Case=Nom|Gender=Masc..."
      },
      "correct_feats": {
        "upos": "NOUN",
        "feats": "Case=Acc|Gender=Masc..."
      },
      "occurrence": 42
    },
    /*
    ... other substitution or insertion
    errors with the same UPOS kernel
     */
  ],
  "['NOUN', '%', 'VERB']": [
    {
      "type": "D",
      "kernel_upos": ["NOUN", "%", "VERB"],
      "kernel_feats": [
        {
          "upos": "NOUN",
          "feats": "Case=Acc..."
        },
        null,
        {
          "upos": "VERB",
          "feats": "Voice=Act..."
        }
      ],
      "deleted_words": ["ے"],
      "occurrence": 15
    }
  ]
}
```


Figure 2: Example of the Final Annotation Data Structure showing a Substitution object (top) and a Deletion object (bottom).

E Infliction Mechanism Details

E.1 Lemma Dictionary Construction

To enable the **Lemma Constraint** described in Section 3.1.3, we required a mapping that could answer the question: *"For a given root lemma, what are all the valid morphological surface forms that exist in standard Urdu?"*

We constructed this knowledge base by iterating through every sentence in the **UD 2.12 Urdu Treebank** (Bhat et al., 2017; Palmer et al., 2009). Since this treebank is manually annotated by linguists, it serves as a ground-truth source for valid morphological variation. The construction process was as

follows:

1. We iterated through all sentences in the treebank.

2. For every token, we extracted its **Lemma**, its **Surface Form**, and its **Morphological Features** (e.g., Case, Gender, Number).

3. We populated two dictionaries; One which stores all surface forms of a lemma, and the second where the key is the *word* which can be a lemma, or a surface form, and the value is a list of observed *Feature Set*.

Figure 3 visualizes the first dictionary structure. Figure 4 shows the second diagram. We use their combination in our infliction algorithm to dynamically generate the ”incorrect” version of a word found in the clean corpus. For example, if the clean corpus contains the masculine verb form *karta* (lemma: *kar*), and the error pattern dictates a switch to feminine, we query the first dictionary to get the key *kar* and query the second dictionary to retrieve the valid feminine form *karti*, provided it was observed in the treebank with the required feature set.

```
{
    "لانا": [
        "لائیں", "لائے", "لانے", "لایا",
        "لائی", "لانا",  "لاتے", "لائیں_گے",

    ]
}
```

Figure 3: Structure of the pre-compiled lemma-to-word dictionary. The lemma serves as the key, pointing to a list of all its derived forms observed in the treebank.


```
{
    "ہیں": [
        {
            "id": "35",
            "text": "ہیں",
            "lemma": "ہے",
            "upos": "VERB",
            "xpos": "VM",
            "feats": "Mood=Ind|Number=Plur|
            Person=3|Tense=Pres|VerbForm=Fin|
            Voice=Act",
            "head": "9",
            "deprel": "acl:relcl",
            "deps": "_",
            "misc": "SpaceAfter=No|Vib=ہے|
            Tam=hE|ChunkId=VGF2|ChunkType=head|
            Stype=declarative|Translit=hīñ|
            LTranslit=he"
        },
        {
            "id": "15",
            "text": "ہیں",
            "lemma": "ہے",
            "upos": "AUX",
            "xpos": "VAUX",
            "feats": "Gender=Masc|Mood=Ind|
            Number=Plur|Person=3|Tense=Pres|
            VerbForm=Fin",
            "head": "13",
            "deprel": "aux",
            "deps": "_",
            "misc": "SpaceAfter=No|Vib=ہے|
            Tam=hE|ChunkId=VGF|ChunkType=child|
            Translit=hīñ|LTranslit=he"
        },
        // other occurrences of "ہیں" continue
        here
    ]
}
```


Figure 4: Example of the UD 2.12 treebank data represented as a dictionary. This is used during the Candidate Search phase to ensure that a target incorrect form corresponds to a valid grammatical usage.

E.2 Insertion and Deletion Context Validation

While Substitutions rely on lemma lookups, Insertion and Deletion infliction relies on strict context matching. To ensure high fidelity:

- **Pre-processing:** We apply character normalization to the clean text and run the Stanza pipeline to generate UPOSFeats objects.
- **Sliding Window:** We utilize a sliding window of size k (set to $k = 3$ for our primary experiments). To optimize performance, we compute a hash of the k-gram UPOS tag sequence to instantly retrieve relevant error patterns from the Annotations Database.
- **OOV Safety:** Before any deletion or insertion is finalized, we perform an Out-of-Vocabulary

(OOV) check on the target word. This prevents the system from learning or propagating typos or rare spellings that may have existed in the noisy Wikipedia source.

F Training Hyperparameters

All neural models in this work were trained using a consistent set of standard hyperparameters to ensure fair comparison across experiments. These values were selected based on preliminary validation experiments and computational constraints.

Hyperparameter	Value
Maximum sequence length	128
Batch size	8
Gradient accumulation steps	4
Effective batch size	32
Optimizer	Adafactor
Learning rate	5×10^{-6}
Number of epochs	1
Mixed precision	bfloat16 (bf16)
Prompt (prefixed to input)	"Correct the grammar of the following Urdu sentence: "

Table 9: Training hyperparameters used for all model fine-tuning experiments.

G Random Noise Implementation

Our random noise baseline follows the statistical noise injection method described by Grundkiewicz et al. (2019). We applied word-level and character-level noise to clean sentences with the following parameters:

Error Rates: For each sentence, an error probability p was sampled from a normal distribution $\mathcal{N}(\mu, \sigma)$. We set $\mu = 0.2$ based on the empirical error rate observed in UrduGEC-Gold, and $\sigma = 0.2$ following the reference study.

Operations: Edits were applied based on the operation distribution: Substitution (0.7), Deletion (0.1), Insertion (0.1), and Swap (0.1). Furthermore, character-level noise (typos) was applied to 10% of tokens.

Confusion Sets: The reference study utilizes Aspell to generate confusion sets for word substitutions. Due to the lack of a robust phonetic spellchecker for Urdu, we generated confusion sets using Levenshtein Edit Distance. For any given word, candidates were selected from the vocabulary if they were within an edit distance of 2. In cases where no Levenshtein neighbors were found, the system fell back to random word substitution.

H Macro Error Categories

For analysis and reporting, fine-grained ERRANT tags are grouped into linguistically motivated macro categories. Each macro category aggregates all operation types—Missing (M), Unnecessary (U), and Replacement (R)—for the listed tags.

H.1 Verb Auxiliaries

This category covers the verbal complex, including tense, aspect, mood, and subject–verb agreement. It includes both root substitutions and inflectional changes affecting main verbs and auxiliary verbs.

ERRANT tags included: VERB, VERB:INFL, AUX, AUX:INFL

H.2 Nouns Pronouns

This category encompasses nominal entities and captures errors related to gender, number, and case, particularly the oblique form required before postpositions. Errors involving proper nouns are also included.

ERRANT tags included: NOUN, NOUN:INFL, PRON, PRON:INFL, PROPN, PROPN:INFL

H.3 Adpositions

This category focuses on the Urdu adpositional system, including case markers such as *ne* and *ko*, as well as genitive agreement forms *ka*, *ke*, and *ki*.

ERRANT tags included: ADP, ADP:INFL

H.4 Modifiers Miscellaneous

A catch-all category that includes modifiers and function words, such as adjectives, adverbs, determiners, conjunctions, particles, and numerals.

ERRANT tags included: ADJ, ADJ:INFL, ADV, DET, CCONJ, SCONJ, PART, NUM, X

H.5 Orthography

This category represents non-grammatical surface-level errors, including spelling mistakes, typographical errors, and incorrect punctuation placement.

ERRANT tags included: SPELL, PUNCT

I LLM-as-a-Judge Evaluation Prompts

To evaluate the quality of the generated corrections in Section 5.4, we utilized five distinct prompt configurations inspired by Kobayashi et al. (2024). To conserve space, we present the common prompt template followed by the specific task instructions and scoring scales injected into the template for each evaluation criterion. All prompts were submitted to Gemini 3 Flash (Google DeepMind, 2025) (gemini-3-flash-preview) via the Google Genera-

tive AI API with temperature set to 0 for deterministic outputs; all other parameters were left at their defaults.

I.1 Common Prompt Template

The following system prompt was provided to the judge model. Dynamic variables are denoted in brackets, where {incorrect_text} is the source sentence, {targets_str} is a list of model predictions, and {output_format} defines the expected JSON schema.

The goal of this task is to rank the presented targets based on the quality of the sentences.
The context is the original sentence written by an Urdu learner.
The targets are predicted versions of that sentence by multiple models.
After reading the predicted sentence, please assign a score from a minimum of 1 point to a maximum of 5 points to each target based on the quality of the sentence (note that you can assign the same score multiple times).

[CRITERIA-SPECIFIC INSTRUCTIONS AND SCALE]

original sentence
{incorrect_text}

targets
{targets_str}

output format
Return this JSON array filling in the scores and reasoning values for each predicted text:
{output_format}

I.2 Criteria-Specific Instructions

The [CRITERIA-SPECIFIC INSTRUCTIONS AND SCALE] section of the template was replaced with one of the following blocks depending on the evaluation dimension:

1. Grammaticality **Task:** Rate **ONLY the grammaticality** of the target sentence. Ignore how naturally it flows (fluency), as long as it is grammatically valid. Ignore whether it perfectly preserves the original meaning.

- **1:** Completely ungrammatical with severe structural errors.
- **2:** Multiple major grammatical errors that impede understanding.
- **3:** Noticeable grammatical errors, but the core structure is intact.
- **4:** Mostly grammatical with only minor technical errors.
- **5:** Perfect grammatical structure with no errors.

2. Meaning Preservation **Task:** Rate **ONLY how well the target preserves the original meaning** of the context. Ignore minor grammatical or fluency issues in the target sentence, as long as the meaning is conveyed.

- **1:** Completely changes or loses the original meaning.
- **2:** Significant portions of the original meaning are lost or altered.
- **3:** Captures the general idea, but misses important nuances or details.
- **4:** Preserves almost all meaning with very minor deviations.
- **5:** Perfectly preserves the original intent and meaning.

3. Fluency **Task:** Rate **ONLY the fluency** of the target sentence. Ignore minor spelling mistakes, wording choices, or whether the meaning perfectly matches the learner sentence, as long as it sounds natural.

- **1:** Completely unnatural or broken Urdu.
- **2:** Very awkward and difficult to read naturally.
- **3:** Understandable but noticeably unnatural.
- **4:** Mostly natural with minor awkwardness.
- **5:** Completely natural and fluent Urdu.

4. Edit Difficulty **Task:** Rate the **difficulty of the corrections** required to transition from the context to the target. First, implicitly identify the edits made. Then, assign a score based on how complex those edits were to identify and execute. *(Note: If no errors existed and no edits were made, award 5. If an error existed but no edits were made, award 1).*

- **1:** Trivial edits (e.g., basic spelling fixes, simple punctuation).
- **2:** Simple edits (e.g., correcting basic gender/number agreement).
- **3:** Moderate edits (e.g., fixing verb tense conjugation).
- **4:** Complex edits (e.g., restructuring clauses, fixing deep syntax).
- **5:** Highly advanced edits (e.g., complete semantic restructuring).

5. Edit Impact **Task:** Rate the **impact** that the specific edits had on improving the sentence. Assign a score based on how much those specific edits improved the overall quality, clarity, and correctness. *(Note: Same no-edit logic applies as above).*

- **1:** Negative/zero impact (made it worse or missed a glaring error).
- **2:** Minimal impact (fixed trivial issues, major errors remain).
- **3:** Moderate impact (improved sentence, but could be noticeably better).
- **4:** High impact (resolved most significant issues).
- **5:** Transformative impact (perfectly resolved all issues).