HiLearners: Non-Native Spoken Hindi Error Correction

Sourava Kumar Behera[1] **and Rohit Saluja**[1,2]

[1]School of Computing and Electrical Engineering,
Indian Institute of Technology Mandi, Himachal Pradesh, India
[2]BharatGen Consortium
s24131@students.iitmandi.ac.in, rohit@iitmandi.ac.in

Abstract

While majority of current resources rely on formal text corrections, our work shifts the focus to non-native spoken Hindi error correction, which presents unique challenges due to its rich morphology, complex syntax, and distinct error patterns. To address the scarcity of authentic learner data, we introduce HiLearners, a dataset gathered from 2,500 real non-native Hindi speakers across three linguistic backgrounds (English, Bengali, Dravidian), capturing authentic error patterns including transfer errors, overgeneralization patterns, and contextual agreement issues. Furthermore, to overcome data resource limitations, we develop a methodical synthetic data augmentation technique, utilizing Large Language Models (LLMs) with a pattern analysis and controlled generation approach similar to Retrieval-Augmented Generation (RAG), yielding 5,500 carefully verified synthetic examples. Through extensive experiments on individual, mixed, and progressive curriculum-based configurations using multilingual models, we demonstrate that LLM-based synthetic data combined with three-phase curriculum learning significantly boosts performance, achieving a 76.92 GLEU score and surpassing human-only baselines. This work bridges the gap between native-centric error correction research and non-native Hindi learner needs, establishing a realistic assessment standard for advancing low-resource language processing.

1 Introduction

In multilingual contexts, non-native speakers are individuals whose first language (L1) differs from the target language they are acquiring. In our study, non-native Hindi speakers are specifically defined as individuals whose mother tongue is not Hindi encompassing speakers from diverse linguistic backgrounds including English, Bengali, and Dravidian languages who are acquiring Hindi as their other language.

The errors produced by non-native Hindi speakers manifest through distinct systematic patterns that fundamentally differ from both native speaker mistakes and grammatically accurate, intentional multilingual phenomena. Corder (1967) established that non-native speaker errors are systematic deviations occurring when individuals have not yet mastered target language rules, contrasting sharply with occasional performance errors made by native speakers. These systematic L2 errors typically emerge through three primary mechanisms: transfer errors, where non-native speakers apply L1 structures to the target language (Kim, 2025); overgeneralization errors, where Hindi grammatical rules are extended beyond appropriate contexts (Hassan and Rami, 2024); and contextual agreement errors, representing failures to maintain grammatical consistency across complex sentence structures (Rothman and Slabakova, 2018).

Selinker (1987) characterized this systematic nature through interlanguage theory, demonstrating that non-native speakers develop transitional linguistic systems influenced by L1 transfer, learning strategies, and overgeneralization patterns. This framework helps distinguish systematic non-native speaker errors (including those where code-mixing results in grammatical violations) from intentional code-mixing behaviors commonly observed in multilingual Indian communities, where speakers deliberately alternate between languages for communicative purposes without necessarily making grammatical violations.

Hindi's morphologically rich nature, with its intricate gender-number-case agreement system and complex verbal morphology, presents particular challenges for non-native speakers from typologically diverse backgrounds. While Patel et al. (2024) provided evidence of bidirectional transfer effects in Hindi-English contexts, and recent studies show that transfer patterns vary significantly based on L1 backgrounds (Rothman, 2015), spoken Hindi error

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1276–1288

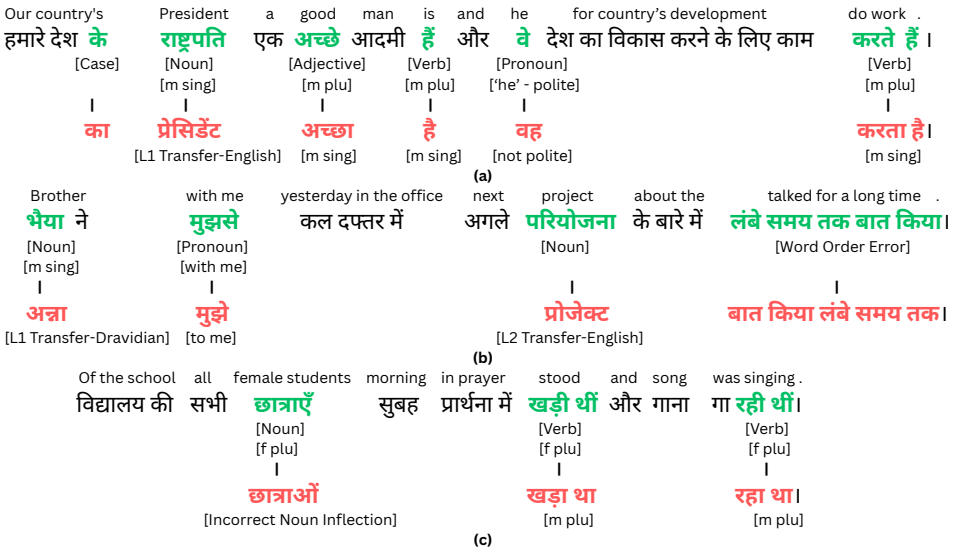

Figure 1: This figure illustrates typical errors found in the HiLearners dataset from Hindi language learners. (For clarity, correct words are highlighted in green, and incorrect words are highlighted in red.) (m:masculine, plu:plural, sing:singular): (a) Errors made by an **English L1 learner**, specifically using English words directly and failing to use the correct polite (honorific) forms for high-status individuals. (b) Examples of multiple error types: influence from **Dravidian (as L1)**, **English lexical interference (as L2)**, incorrect pronoun usage, and mistakes in sentence word order. (c) Errors primarily reflecting **Bengali L1 influence**, particularly concerning verb agreement (gender and number) and noun case/number inflection.

correction remains severely understudied. Most research has focused on formal text correction (Sonawane et al., 2020; Sharma and Bhattacharyya, 2025), leaving a significant gap in understanding authentic non-native speaker errors in spoken production.

Our work specifically targets systematic errors made by non-native speakers, including those arising from code-mixing as a legitimate communicative strategy, focusing on individuals who possess basic Hindi speaking ability and foundational grammar knowledge but lack proficiency in complex grammatical structures and complete mastery of Hindi's intricate morphosyntactic systems (Li et al., 2025).

Thus, we present our contributions as follows:

1. **HiLearners Dataset:** We introduce the first spoken Hindi error correction dataset comprising 2,500 sentences with 24 distinct error types collected from non-native Hindi speakers across three linguistic backgrounds (English, Bengali, Dravidian). The dataset captures systematic transfer errors, overgeneralization patterns, and contextual agreement violations as illustrated in Figure 1. The dataset is available for access at link[1].

2. **Synthetic Data Augmentation:** We develop 5,500 LLM-generated synthetic sentences that authentically replicate learner error patterns. Each synthetic example is verified by native speakers to ensure linguistic validity and maintains error distribution consistency with real non-native speaker production.

3. **Comprehensive Evaluation Framework:** We design a systematic experimental framework employing curriculum learning with three progressive difficulty phases and varying synthetic data proportions (25%-100%) to identify optimal training strategies for spoken Hindi error correction systems.

This work bridges the gap between native-centric NLP research and non-native Hindi speakers' needs. It presents the first systematic collection of authentic spoken Hindi errors from non-native speakers, establishing a benchmark for error correction research and adaptive language learning technologies.

2 Related Works

Non-native speaker errors in second language acquisition are systematic deviations that occur when learners transfer linguistic patterns from their first

[1] https://github.com/Souravakb24/HiLearners

language to the target language. These errors, which Corder (1967) identified as evidence of interlanguage development rather than random mistakes, predominantly manifest as grammatical violations involving gender agreement, case marking, number concordance, and verbal morphology, especially in morphologically rich languages (Lee and Seneff, 2008; Han et al., 2006). Research indicates that speakers from diverse linguistic backgrounds produce distinct error patterns when acquiring the same target language, reflecting the impact of L1 transfer on morphosyntactic structures.

The availability of authentic non-native speaker data has been vital for advancing error correction research. English benefits from extensive learner corpora such as ICLE (Granger et al., 2020), NUCLE (Dahlmeier et al., 2013), TOEFL11 (Blanchard et al., 2013), and crowd-sourced Lang-8 data (Mizumoto et al., 2011; Tajiri et al., 2012). Similar resources exist for other languages, including CEDEL2 for Spanish L2 (Lozano et al., 2009). More recent English resources include W&I+LOCNESS (Bryant et al., 2019) and the CoNLL-2014 benchmark (Ng et al., 2014). These corpora have facilitated robust error correction systems by capturing systematic learner errors and interlanguage phenomena. In contrast, Hindi significantly lacks authentic non-native speaker data. Existing Hindi Error correction research relies on Wikipedia editorial corrections (Sonawane et al., 2020; Sharma and Bhattacharyya, 2025), which capture formal text improvements instead of genuine learner productions, thus failing to represent transfer errors and morphological overgeneralization patterns typical of authentic L2 acquisition.

Error correction methodologies have significantly evolved. Initially, rule-based systems (Foster and Andersen, 2009) used manually crafted grammatical rules. This was followed by statistical approaches, which introduced probabilistic modeling using corpus frequencies and n-gram language models to flag uncommon linguistic sequences as potential errors (Izumi et al., 2004). Early milestones included spell checking systems focused on orthographic errors before expanding to basic grammatical patterns. The shift to neural approaches fundamentally transformed the field, initially conceptualizing the task as machine translation where erroneous text is mapped to corrected forms (Brockett et al., 2006). This was later advanced through transformer models (Vaswani et al., 2017) and sophisticated architectures like multilayer convolutional encoder-decoder networks (Chollampatt and Ng, 2018) and copy-augmented architectures (Zhao et al., 2019) designed for low-resource grammatical error correction tasks. Recent developments have further incorporated pre-training strategies and unsupervised methods (Grundkiewicz et al., 2019) to enhance performance across diverse error types and linguistic contexts. Additionally, specific works like UTTAM (Jain et al., 2018) and SCMIL (Etoori et al., 2018) have applied probabilistic and deep learning approaches, respectively, to spelling correction in Indic languages, while simple n-gram based models (Singh and Singh, 2019; Kanwar et al., 2017) have been used for "RealWord" error correction.

Evaluation frameworks for error correction have seen significant advancements, introducing standardized metrics like the M2 scorer (Dahlmeier and Ng, 2012) for improved error-level assessment, and GLUE metrics (Napoles et al., 2015) as alternatives to traditional BLEU scoring. Annotation tools such as ERRANT (Bryant et al., 2017; Felice et al., 2016) enable systematic categorization of errors, a method adapted for Hindi by Sonawane et al. (2020) and Sharma and Bhattacharyya (2025). However, these existing datasets still exhibit limited coverage of context-based errors commonly encountered by Hindi learners. Our HiLearners dataset addresses this gap, providing a benchmark that captures authentic learner errors for more comprehensive Hindi Error Correction evaluation.

3 Data

This section details the datasets employed in our study: HiLearners, a meticulously human-annotated corpus of non-native Hindi, and a Synthetic Data set specifically designed to augment and complement it.

3.1 HiLearners

The HiLearners dataset is a collection of 2,500 sentences, all generated by non native Hindi speakers. These learners come from three different linguistic backgrounds: English, Bengali, and Dravidian, as illustrated in Figure 2. To gather this data, we designed structured writing tasks specifically to highlight authentic learner errors across various proficiency levels.

To ensure annotation quality, all 2,500 sentences were independently reviewed by three native Hindi linguists who classified errors. We evaluated inter-

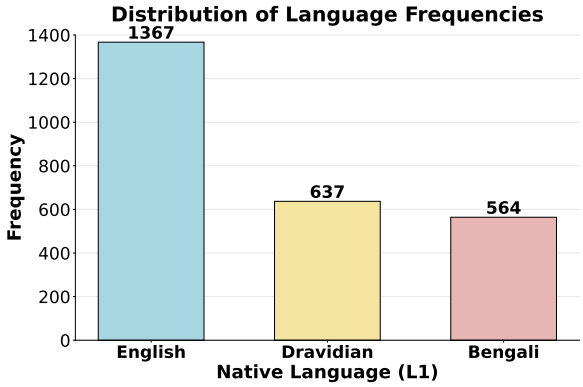

Figure 2: Distribution of sentences by non-native speakers language background (L1).

annotator agreement using Cohen's Kappa coefficient (Cohen, 1960), achieving a score of 0.96 at the sentence level, which happens because the annotators are native Hindi speakers. The HiLearners sentences contain anywhere from one to six errors per sentence.

Our focus in HiLearners is on three common types of errors encountered in language acquisition:

1. **Transfer Errors:** These involve issues with morphology and syntax that arise from interference from the learner's first language (L1).

2. **Overgeneralization Errors:** This category includes the incorrect application of Hindi grammatical rules, such as gender-number agreement. This is particularly common among speakers whose native languages don't have grammatical gender.

3. **Contextual Agreement Errors:** These errors highlight difficulties in maintaining grammatical consistency within complex sentence structures, indicating struggles with long-range dependencies.

It's important to note that most errors found in morphology, pronouns, and nouns are typically categorized as either L1/L2 transfer errors or overgeneralization errors. Other error types generally encompass overgeneralization and contextual agreement errors.

A significant portion of the errors within HiLearners include issues with word order, phrase order, and transliteration. Sentences with more than two errors were primarily manually annotated by our linguists. This was crucial because certain error types, such as transliteration and word order, simply cannot be handled by automated tools such as ERRANT. Through this comprehensive process, we successfully identified and categorized a total of 24 distinct error types. You can see examples of these in Figure 3, and Figure 4 further illustrates the frequency distribution of sentences based on the number of annotated errors.

3.2 Synthetic Data Generation

To address data resource limitations, we developed a synthetic data augmentation technique generating 5,500 sentences using ChatGPT from the training and validation splits of HiLearners. Our RAG-inspired approach employed pattern analysis and controlled generation in a three-step methodology to replicate authentic error patterns from the training data.

1. **Pattern Analysis**: We started by conducting a thorough analysis of the human-annotated HiLearners data. This involved providing ChatGPT and Claude with pairs of incorrect and corrected sentences, allowing it to identify and extract the underlying error patterns across all defined categories. This step gave LLM a clear understanding of how these errors manifest and what types of changes are typically made during correction. This process mirrors the "retrieval" aspect of RAG, where the model learns from existing knowledge.

2. **Controlled Generation**: We utilized ChatGPT and Claude and fed it clean Hindi sentences from the BPCC corpus (Gala et al., 2023) and specifically prompted it to introduce errors that replicated the patterns identified in the HiLearners dataset. This "generation" phase, informed by the learned patterns, allowed us to create new examples with controlled error types.

3. **Native Speaker Verification**: Finally, every synthetically generated sentence was subjected to rigorous manual verification by native Hindi speakers. This crucial step ensured the linguistic authenticity of the sentences and confirmed that the induced errors genuinely reflected real learner mistakes. During this verification, we also ensured that the filtered sentences contained errors ranging from 1 to 6 per sentence, consistent with the distribution observed in our Hilearners data. From the total output, we carefully selected the 5,500

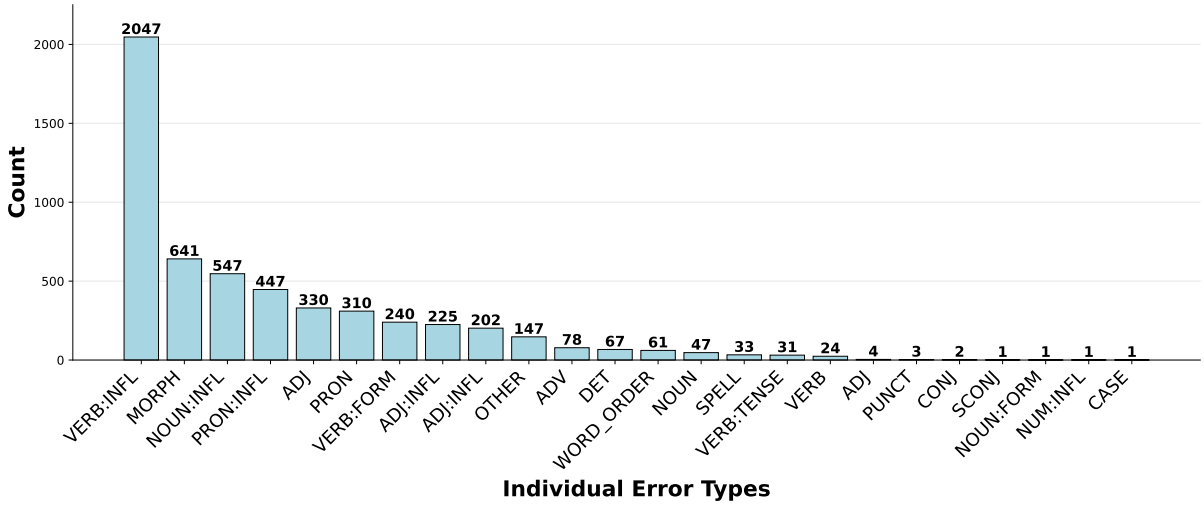

Figure 3: Examples of the types of errors in HiLearners

Data	#Sentence Pairs	#Tokens
Human	2500	113336
Synthetic	5500	360412

Table 1: Dataset Statistics with Number of sentence pairs and tokens

sentences that accurately mirrored the error characteristics of the HiLearners dataset.

This synthetic dataset is stratified by error count and type, enabling curriculum learning where models progressively learn from simple to complex error patterns, optimizing correction capabilities.

Table 1 provides a detailed overview of the statistics for both the human-annotated HiLearners and the synthetically generated datasets, including the number of sentence pairs and tokens.

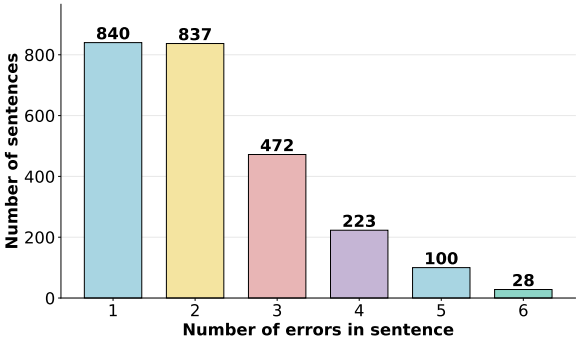

Figure 4: Distribution of sentences based on the number of annotated errors.

4 Experiments

We conduct comprehensive experiments to evaluate the effectiveness of our hybrid dataset approach for Hindi Error Correction. Our experimental framework systematically investigates optimal data composition strategies through varying synthetic data proportions and implementing curriculum learning with progressive difficulty phases.

4.1 Models

We evaluate three multilingual language models that have demonstrated strong performance on low-resource language tasks inclusing Hindi. IndicBART (Dabre et al., 2021) is a sequence-to-sequence model specifically designed for Indian languages, pre-trained on 11 Indian languages including Hindi using the ai4bharat/IndicBART checkpoint. mT5-large (Xue et al., 2020) represents a multilingual variant of T5 trained on the mC4 corpus covering 101 languages, employed through the google/mt5-large checkpoint. mBART-large-50 (Tang et al., 2020) is a multilingual denoising auto-encoder pre-trained on 50 languages using the facebook/mbart-large-50 checkpoint, which has shown effectiveness in multilingual text generation tasks.

4.2 Dataset Mixing Strategy

Our data mixing experiments systematically evaluate the impact of synthetic data augmentation by training models with varying proportions of synthetic data combined with human-annotated data. The mixing strategy includes configurations with 25%, 50%, 75%, and 100% synthetic data propor-

tions, alongside pure human-only and synthetic-only baselines. Each mixed configuration maintains the complete human-annotated dataset (1,750 samples) while incrementally adding synthetic data ranging from 954 samples (25%) to 3,822 samples (100%) to analyze the optimal balance between authentic learner errors and scaled synthetic patterns. All synthetic data portions employ stratification by error count to ensure balanced representation across different error complexities, enabling systematic analysis of how synthetic data quantity affects model performance on authentic learner errors.

4.3 Curriculum Learning

We implemented a progressive three-phase curriculum learning approach, mimicking natural language acquisition by gradually increasing error complexity. Each phase utilizes a carefully curated mix of human and synthetic data:

1. **Easy Phase**: Targets simpler 1-2 error sentences with synthetic samples, building foundational correction capabilities.

2. **Medium Phase**: Introduces moderate complexity (1-4 errors), including agreement and syntactic issues, expanding the model's exposure through additional synthetic data.

3. **Hard Phase**: Encompasses the full range of error densities (1-6 errors), including complex multi-error scenarios, leveraging all available synthetic data to prepare the model for real-world complexities.

This phased approach allows models to develop foundational correction skills before tackling more challenging error patterns. Specific data splits for all experiments are detailed in Appendix B.

4.4 Training and Evaluation Setup

All models undergo standard sequence-to-sequence fine-tuning using cross-entropy loss with detailed hyperparameters provided in Appendix A. We employ early stopping with patience of 5 epochs based on validation performance to prevent overfitting. Each experiment maintains consistent evaluation using 375 human-annotated test samples from HiLearners and 375 error-free test samples to ensure fair comparison across different training configurations.

We evaluate model performance using multiple complementary metrics. GLEU (Napoles et al., 2015) serves as our primary evaluation metric given its strong correlation with human judgment for Error Correction tasks. We supplement this with $F_{0.5}$ scores computed using token-level alignment following established Error Correction evaluation protocols (Dahlmeier and Ng, 2012).

5 Results

We evaluated three state-of-the-art multilingual models across different data mixing strategies and curriculum learning phases, revealing optimal training configurations for Hindi Error Correction.

5.1 Model Performance Analysis

IndicBART, despite its Indian language focus, shows modest but consistent improvements. The model achieves gains of $+2.67$ GLEU and $+0.017$ $F_{0.5}$ from human-only to optimal configuration, with stable performance across various settings indicating reliable but limited correction capabilities. **mT5-large** demonstrates the most dramatic improvement trajectory, showing exceptional sensitivity to synthetic data augmentation. The model progresses from baseline performance (42.74 GLEU, 0.213 $F_{0.5}$) to peak performance (63.98 GLEU, 0.519 $F_{0.5}$), representing gains of $+21.24$ GLEU and $+0.306$ $F_{0.5}$. **mBART-large-50** achieves superior performance across all configurations, reaching peak scores of 76.92 GLEU and 0.693 $F_{0.5}$ in the optimal setting. While its GLEU score shows a modest increase (e.g., $+1.94$ GLEU from Human-Only to Mixed (100%) or Hard Phase), the $F_{0.5}$ score demonstrates a more significant improvement (e.g., $+0.135$ $F_{0.5}$ from Human-Only to Mixed (100%) or Hard Phase). This pattern of a notable $F_{0.5}$ gain with relatively little GLEU inflation is consistently observed across both synthetic data mixing strategies and curriculum learning phases. The model's multilingual denoising pre-training provides robust error correction capabilities, maintaining strong precision-recall balance and achieving high $F_{0.5}$ scores, with most configurations exceeding 0.65.

5.2 Synthetic Data Mixing Analysis

Our systematic evaluation reveals that progressive synthetic data inclusion directly correlates with improved performance across all models. While synthetic-only configurations underperform significantly (mT5-large: -11.88 GLEU, -0.139 $F_{0.5}$ vs. human-only), the Mixed (100%) configuration

Experiment	IndicBART				mT5-large				mBART-large-50			
	GLEU	P	R	$F_{0.5}$	GLEU	P	R	$F_{0.5}$	GLEU	P	R	$F_{0.5}$
Human-Only	46.26	0.237	0.275	0.237	42.74	0.212	0.249	0.213	74.98	0.583	0.538	0.558
Synthetic-Only	37.48	0.112	0.128	0.112	30.86	0.083	0.074	0.074	47.45	0.357	0.260	0.310
Mixed (25%)	47.53	0.239	0.283	0.240	42.59	0.195	0.226	0.195	74.31	0.680	0.640	0.655
Mixed (50%)	47.85	0.241	0.284	0.242	55.07	0.395	0.377	0.370	75.74	0.688	0.625	0.656
Mixed (75%)	48.32	0.248	0.290	0.249	61.09	0.525	0.425	0.481	75.68	0.692	0.637	0.662
Mixed (100%)	**48.62**	**0.252**	**0.296**	**0.253**	**62.18**	**0.534**	**0.432**	**0.487**	**76.24**	**0.701**	**0.665**	**0.676**

Table 2: Performance comparison across multilingual models on Hindi error correction showing GLEU scores, Precision (P), Recall (R), and $F_{0.5}$ scores for base models (Human only & Synthetic only) and different data mixing strategies. Best model is represented in bold

Experiment	IndicBART				mT5-large				mBART-large-50			
	GLEU	P	R	$F_{0.5}$	GLEU	P	R	$F_{0.5}$	GLEU	P	R	$F_{0.5}$
Easy Phase	48.03	0.240	0.284	0.242	57.22	0.415	0.391	0.390	76.22	0.704	0.656	0.677
Medium Phase	48.55	0.251	0.291	0.251	62.19	0.539	0.441	0.497	76.74	0.715	0.653	0.684
Hard Phase	**48.93**	**0.253**	**0.299**	**0.254**	**63.98**	**0.559**	**0.465**	**0.519**	**76.92**	**0.718**	**0.672**	**0.693**

Table 3: Performance comparison across multilingual models on Hindi error correction showing GLEU scores, Precision (P), Recall (R), and $F_{0.5}$ scores for different curriculum learning phases. Best model is represented in bold

consistently achieves optimal results, demonstrating the power of knowledge distillation from large language models (Table 2). Key findings include:

1. **Human-only baselines are insufficient** – all models only reach their best performance when either 100% synthetic data augmentation or Hard Phase data is incorporated.

2. **Synthetic data consistently boosts performance** – Across all models, we observe a significant and consistent increase in scores as synthetic data is progressively added during mixed-data training.

5.3 Curriculum Learning Analysis

The three-phase curriculum approach demonstrates progressive performance gains. mT5-large shows clear progression: Easy Phase (57.22 GLEU, 0.390 $F_{0.5}$) $\rightarrow$ Medium Phase ($+4.97$ GLEU, $+0.107$ $F_{0.5}$) $\rightarrow$ Hard Phase ($+1.79$ GLEU, $+0.022$ $F_{0.5}$), with Hard Phase consistently matching Mixed (100%) results across all models (Table 3). mBART-large-50 achieves the best results in low-resource scenarios.

Table 4 shows curriculum learning's impact on complex error handling with mBART-large-50. Easy Phase excels for 1-2 errors, while Hard Phase demonstrates superior robustness for higher error counts: 3-error sentences show $+2.45$ GLEU and $+0.026$ $F_{0.5}$ over Medium Phase; 6-error sentences achieve $+7.40$ GLEU and $+0.039$ $F_{0.5}$ improvement over Easy Phase. Hard Phase's exposure to diverse synthetic patterns enables better handling of challenging multi-error scenarios, validating our curriculum design. Figure 5 illustrates model performance by correction status; detailed error analysis is in Appendix C.

5.4 Error-free test samples Analysis

Table 5 showing three multilingual models on the test set consisting of fully correct samples. This evaluation is crucial as it measures the model's ability to avoid making spurious corrections to error-free Hindi text.

The near perfect scores, especially for mBART-large-50, confirm that the training methodology does not significantly compromise the models' ability to preserve fluency and grammaticality when no error correction is needed. This is an important validation of the model's overall robustness.

Model	GLEU	$F_{0.5}$
IndicBART	96.01	0.86
mT5-large	98.09	0.91
mBART-large-50	**99.19**	**0.97**

Table 5: Performance Metrics on the test set of fully correct samples showing GLEU scores, Precision (P), Recall (R), and $F_{0.5}$ scores for all models

#Errors	Easy Phase				Medium Phase				Hard Phase			
	GLEU	P	R	$F_{0.5}$	GLEU	P	R	$F_{0.5}$	GLEU	P	R	$F_{0.5}$
1	**87.09**	**0.63**	**0.68**	**0.64**	86.03	0.629	0.668	0.628	86.44	0.6140	0.6507	0.6141
2	**81.15**	**0.74**	**0.65**	**0.70**	78.04	0.687	0.635	0.662	78.64	0.676	0.631	0.652
3	80.97	0.76	0.69	0.73	78.63	0.754	0.671	0.721	**81.08**	**0.777**	**0.699**	**0.747**
4	65.15	0.74	0.55	0.66	**68.66**	**0.766**	**0.608**	**0.704**	66.97	0.705	0.558	0.640
5	59.34	0.66	0.48	0.58	58.43	0.612	0.486	0.594	**60.66**	**0.681**	**0.518**	**0.675**
6	57.79	0.66	0.39	0.54	57.79	0.630	0.392	0.547	**65.19**	**0.667**	**0.487**	**0.586**

Table 4: Performance analysis by error count per sentence comparing Easy Phase, Medium Phase, and Hard Phase models using mBART-large-50 with best model for each error count is represented by bold.

Incorrect	Correct	Response	Status
मेरी तीन घड़ियों में से एक बंद हो **गया** है और मरम्मत करानी पड़ेगी। (**Merī** tīn ghaṛiyoṅ meṅ se ek band ho **gayā** hai aur marammat karānī paṛegī.)	**मेरी** तीन घड़ियों में से एक बंद हो **गई** है और मरम्मत करानी पड़ेगी। (**Merī** tīn ghaṛiyoṅ meṅ se ek band ho **gaī** hai aur marammat karānī paṛegī.)	**मेरी** तीन घड़ियों में से एक बंद हो **गई** है और मरम्मत करानी पड़ेगी। (**Merī** tīn ghaṛiyoṅ meṅ se ek band ho **gaī** hai aur marammat karānī paṛegī.)	correct
वह मुझे फोन **के ऊपर** बुलाया और कहा कि वह देर से आएगा। (**Vah** mujhe fon **ke ūpar** bulāyā aur kahā ki vah der se āegā.)	**उसने** मुझे फोन **पर** बुलाया और कहा कि वह देर से आएगा। (**Usne** mujhe fon **par** bulāyā aur kahā ki vah der se āegā.)	**वह** मुझे फोन **पर** बुलाया और कहा कि वह देर से आएगा। (**Vah** mujhe fon **par** bulāyā aur kahā ki vah der se āegā.)	partially correct
मैंने हिंदी फिल्म **देखता** था जब मेरा दोस्त मुझे मिलने आया था। (**Maine** Hindī philm **dekhtā** thā jab merā dost mujhe milne āyā thā.)	**मैं** हिंदी फिल्म **देख रहा** था जब मेरा दोस्त मुझे मिलने आया था। (**Main** Hindī philm **dekh rahā** thā jab merā dost mujhe milne āyā thā.)	**मैंने** हिंदी फिल्म **देखते** समय मेरा दोस्त मुझे मिलने **आये थे**। (**Maine** Hindī philm **dekhte** samay merā dost mujhe milne **āye the**.)	incorrect
वह कमरे में बैठा और अखबार पढ़ता था **जब मैं पहुँचा**। (Vah kamre meṅ baiṭhā aur akhbār paṛhtā thā **jab maiṁ pahuṅcā**.)	**जब मैं पहुँचा**, वह कमरे में बैठी और अखबार पढ़ रही थी। (**Jab maiṁ pahuṅcā**, vah kamre meṅ baiṭhī aur akhbār paṛh rahī thī.)	वह कमरे में बैठा और अखबार पढ़ता था **जब मैं पहुँचा**। (Vah kamre meṅ baiṭhā aur akhbār paṛhtā thā **jab maiṁ pahuṅcā**.)	unchanged

Figure 5: Examples on mBART-large-50 Hard Phase model Performance according to correction status

5.5 Results on LLMs

The zero-shot results on LLMs such as GPT-5, Gemini 2.5 Flash, and Claude Sonnet 4.5 are shown in Table 6. LLMs have shown relatively mid-level performance scores compared to fine-tuned models.

Model	GLEU	P	R	$F_{0.5}$
Gemini 2.5 Flash	67.23	0.52	0.56	0.51
Claude Sonnet 4.5	69.05	0.58	0.52	0.55
GPT-5	**70.40**	**0.58**	**0.59**	**0.57**

Table 6: Performance metrics on the test set showing GLEU scores, Precision (P), Recall (R), and $F_{0.5}$ scores for all models

6 Conclusion

Several limitations warrant consideration. The HiLearners dataset, while pioneering in focusing on non-native speakers, comprises only 2,500 sentences from three linguistic backgrounds (English, Bengali, Dravidian), which may not fully represent the diversity of Hindi learner populations. Our synthetic augmentation approach, despite native speaker verification, may inadequately capture complex authentic errors arising from cultural and pragmatic factors. The sentence-level evaluation framework does not address discourse coherence or contextual appropriate factors critical for real-world applications. Additionally, our 19-category error taxonomy may not encompass all error patterns across varying proficiency levels, and computational constraints restricted our experiments to three multilingual models, leaving potential architecture-specific optimizations unexplored.

7 Limitations

While our work advances Hindi GEC through authentic learner data and systematic synthetic augmentation, several limitations exist. The HiLearners dataset (2,500 sentences) covers only three linguistic backgrounds (English, Bengali, Dravidian), potentially limiting generalizability to broader Hindi learner populations. Our synthetic data generation, though verified by native speakers, may not fully capture nuanced authentic learner errors, particularly those from cultural and pragmatic contexts. The evaluation framework focuses on sentence-level corrections and may inadequately assess discourse-level coherence or contextual appropriateness crucial for practical applications. Our 19-category error taxonomy may overlook emerging patterns specific to different proficiency levels. Computational constraints limited evaluation to three multilingual models, and observed performance gaps suggest unexplored architecture-specific optimizations could yield further improvements.

8 Ethics Statement

The HiLearners dataset was collected by a university Hindi linguist from non-native speakers across diverse linguistic backgrounds, with explicit participant consent for research use. All contributors participated voluntarily with full knowledge of research objectives. No personally identifiable or sensitive information was collected; all data is anonymized and aggregated to ensure privacy protection. Participants were informed that their language samples would contribute to grammatical error correction systems for educational and accessibility purposes. This work adheres to institutional ethical guidelines for linguistic data collection and computational research.

9 Acknowledgments

This work is supported by and is part of BharatGen[2], an Indian Government-funded initiative focused on developing multimodal large language models for Indian languages. We thank Manikandan Ravikiran for his insightful initial reviews.

References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Chris Brockett, Bill Dolan, and Michael Gamon. 2006. Correcting esl errors using phrasal smt techniques. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, Australia*.

Christopher Bryant, Mariano Felice, Øistein E Andersen, and Ted Briscoe. 2019. The bea-2019 shared task on grammatical error correction. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 52–75.

CJ Bryant, Mariano Felice, and Edward Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. Association for Computational Linguistics.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Stephen Pit Corder. 1967. The significance of learner's errors.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. 2021. Indicbart: A pre-trained model for indic natural language generation. *arXiv preprint arXiv:2109.02903*.

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.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 22–31.

Pravallika Etoori, Manoj Chinnakotla, and Radhika Mamidi. 2018. Automatic spelling correction for resource-scarce languages using deep learning. In *Proceedings of ACL 2018, Student Research Workshop*, pages 146–152.

Mariano Felice, Christopher Bryant, and Ted Briscoe. 2016. Automatic extraction of learner errors in esl sentences using linguistically enhanced alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 825–835.

[2] https://bharatgen.tech/

Jennifer Foster and Øistein E Andersen. 2009. Generrate: Generating errors for use in grammatical error detection. The Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, and 1 others. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Sylviane Granger, Maïté Dupont, Fanny Meunier, Hubert Naets, and Magali Paquot. 2020. *International Corpus of Learner English. Version 3*.

Roman Grundkiewicz, Marcin Junczys-Dowmuntz, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *14th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263. Association for Computational Linguistics.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in english article usage by non-native speakers. *Natural Language Engineering*, 12(2):115–129.

DMK Hassan and Zawad Rami. 2024. Exploring nuances in second language acquisition: an error analysis perspective. *Migration Letters*, 21(4):294–298.

Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2004. Sst speech corpus of japanese learners' english and automatic detection of learners' errors. *ICAME journal*, 28:31–48.

Amita Jain, Minni Jain, Goonjan Jain, and Devendra K Tayal. 2018. "uttam" an efficient spelling correction system for hindi language based on supervised learning. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(1):1–26.

Shailza Kanwar, Manoj Sachan, and Gurpreet Singh. 2017. N-grams solution for error detection and correction in hindi language. *International Journal of Advanced Research in Computer Science*, 8(7).

Hyunwoo Kim. 2025. Crosslinguistic influence in bilingual morphosyntactic processing: Effects of language-common, language-contrasting, and language-specific information. *Bilingualism: Language and Cognition*, 28(1):172–184.

John Lee and Stephanie Seneff. 2008. An analysis of grammatical errors in non-native speech in english. In *2008 IEEE spoken language technology workshop*, pages 89–92. IEEE.

Shaofeng Li, Ling Ou, and Icy Lee. 2025. The timing of corrective feedback in second language learning. *Language Teaching*, pages 1–17.

Cristóbal Lozano and 1 others. 2009. Cedel2: Corpus escrito del español l2.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning sns for automated japanese error correction of second language learners. In *Proceedings of 5th international joint conference on natural language processing*, pages 147–155.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The conll-2014 shared task on grammatical error correction. In *Proceedings of the eighteenth conference on computational natural language learning: shared task*, pages 1–14.

Priyanka Patel, Nandini Chatterjee Singh, and Minna Torppa. 2024. Understanding the role of cross-language transfer of phonological awareness in emergent hindi–english biliteracy acquisition. *Reading and Writing*, 37(4):887–920.

Jason Rothman. 2015. Linguistic and cognitive motivations for the typological primacy model (tpm) of third language (l3) transfer: Timing of acquisition and proficiency considered. *Bilingualism: language and cognition*, 18(2):179–190.

Jason Rothman and Roumyana Slabakova. 2018. The generative approach to sla and its place in modern second language studies. *Studies in second language acquisition*, 40(2):417–442.

Larry Selinker. 1987. Some unresolved issues in an elt new media age: Towards building an interlanguage semantics. *Language*, page 995.

Ujjwal Sharma and Pushpak Bhattacharyya. 2025. Hi-GEC: Hindi grammar error correction in low resource scenario. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6063–6075, Abu Dhabi, UAE. Association for Computational Linguistics.

Shashank Singh and Shailendra Singh. 2019. Handling real-word errors of hindi language using n-gram and confusion set. In *2019 Amity International Conference on Artificial Intelligence (AICAI)*, pages 433–438. IEEE.

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.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for esl learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 198–202.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. *arXiv preprint arXiv:1903.00138*.

A Implementation Details

Parameter	Value
Number of Training Epochs	4
Per Device Train Batch Size	6
Per Device Eval Batch Size	6
Gradient Accumulation Steps	3
Warmup Steps	75
Weight Decay	0.01
Learning Rate	2e-5
LR Scheduler Type	"linear"
Max Gradient Norm	1.0
Adam Epsilon	1e-8
Adam Beta1	0.9
Adam Beta2	0.999

Table 7: Training parameters used for all models.

The key training parameters used for all models (IndicBART, mT5-large, and mBART-large-50) in our experiments mentioned in Table 7. These parameters were consistently applied across all data mixing strategies and curriculum learning phases to ensure a fair comparison of model performance under different data compositions. All models were fine-tuned using their official pre-trained versions available on Hugging Face.

B Training Data Splits

Table 8 presents the experimental configurations across different data mixing strategies and curriculum learning phases used in this study. The baseline experiments include a human-only configuration and a synthetic-only setup, both utilizing human-annotated test samples from HiLearners for evaluation. Mixed data experiments progressively combine human and synthetic data at various ratios, with corresponding training and validation sets. Crucially, in these mixed data configurations, the training and validation samples maintain errors relative to their respective dataset contributions (error count per sentence). Furthermore, in mixed data experiments, errors were selected with proper proportion according to their counts. The curriculum learning approach implements three phases with increasing difficulty: Easy Phase, Medium Phase, and Hard Phase. All experimental configurations maintain the same human-annotated data foundation while systematically varying the synthetic data integration to evaluate the impact of data augmentation and curriculum learning strategies on grammatical error correction performance.

C Error Analysis

Figure 6 illustrates the distribution of correction statuses across different difficulty phases. The Easy Phase demonstrates a higher proportion of correctly corrected sentences compared to the Medium and Hard Phases. Conversely, the Hard Phase yields more partially correct sentences than the Medium and Easy Phases. Interestingly, the Easy Phase also shows a greater incidence of incorrect and unchanged sentences. Overall, the Hard Phase appears to achieve the most favorable balance of correction outcomes.

Further analysis of error types, presented in Figure 7, reveals significant performance variations across linguistic categories on the test set. The model achieves over 80% correction rates for systematic grammatical patterns and well-represented linguistic structures. However, morphologically complex errors, particularly those involving noun and pronoun morphology, and context-dependent corrections pose greater challenges, with correction rates falling below 60%. This substantial error rate in morphological categories, especially for nouns and pronouns, suggests a difficulty in recognizing transfer errors potentially influenced by L1/L2 interference. This hierarchy indicates that certain

Experiments	Train Samples	Val Samples	Human Data	Synthetic Data
Human-Only Baseline	1,750	375	1,750	0
Synthetic-Only	3,822	819	0	3,822
Mixed Data (25%)	2,704	579	1,750	954
Mixed Data (50%)	3,659	783	1,750	1,909
Mixed Data (75%)	4,612	986	1,750	2,862
Mixed Data (100%)	5,572	1,194	1,750	3,822
Easy Phase	3,391	727	1,750	1,641
Medium Phase	5,492	1,177	1,750	3,742
Hard Phase	5,572	1,194	1,750	3,822

Table 8: Experimental configurations showing training and validation sample sizes across different data mixing strategies and curriculum learning phases. All experiments use 375 human-annotated test samples from HiLearners for evaluation.

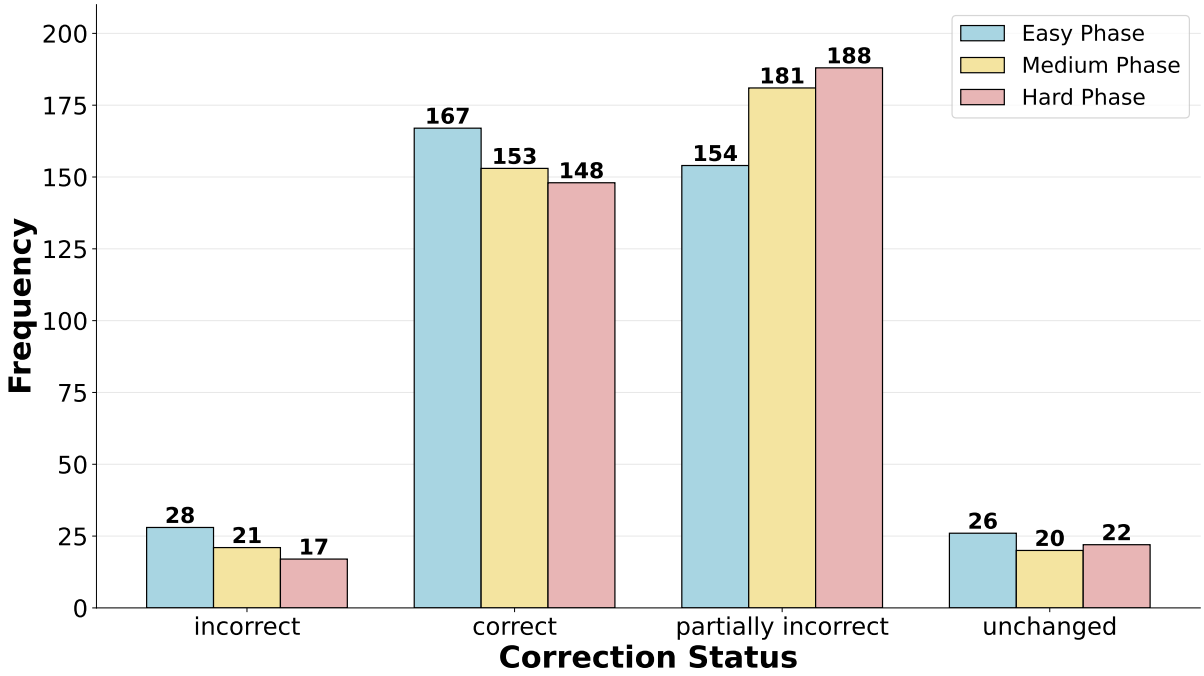

Figure 6: Frequency distribution of correction statuses (incorrect, correct, partially incorrect, and unchanged) across Easy, Medium, and Hard phases in mBART-large-50 model.

error types benefit more from the curriculum learning progression.

The positive impact of curriculum learning is evident in the model's enhanced contextual understanding and improved recognition of systematic error patterns. The Hard Phase results specifically highlight successful adaptation to increasingly difficult linguistic phenomena, particularly in handling multi-error scenarios.

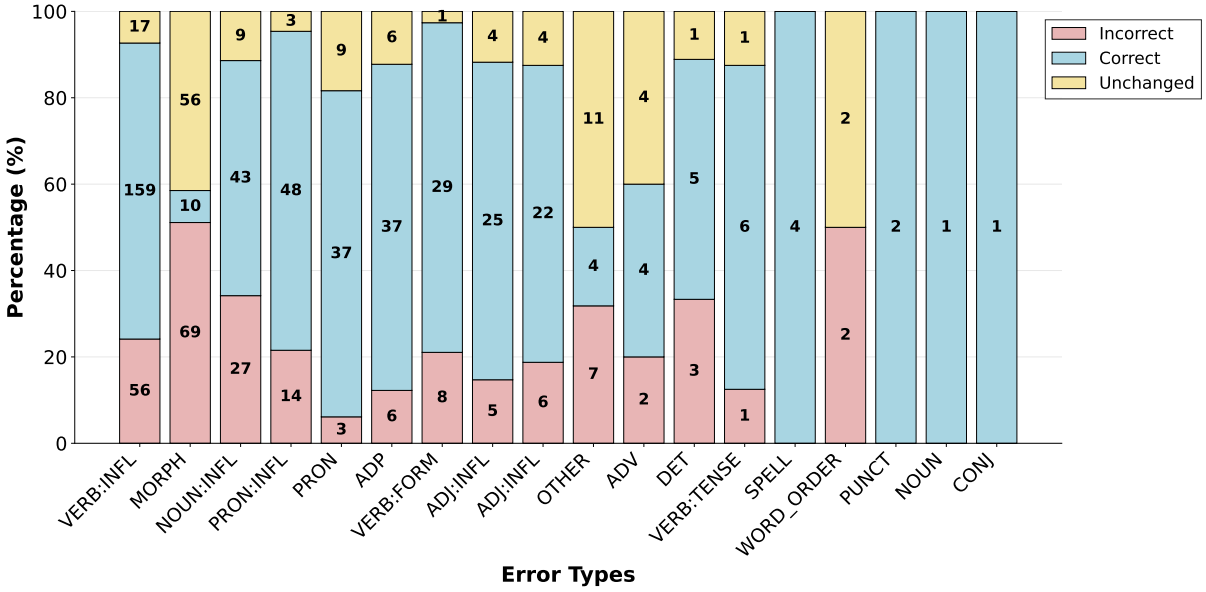

Figure 7: Percentage distribution of error correctness on test set across various error types on mBART-large-50 Hard Phase model.