Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 2, 2026

DEEP LEARNING FOR GRAMMER AND SYNTAX CORRECTION IN ENGLISH LANGUAGE TEACHING

Muhammad Zahid Sarwar[1], Shiraz Hassan[*2]

[1,*2]*Department of Computer Science, CECOS University of IT and Emerging Sciences, Peshawar*

[*2]shirazhassan@cecos.edu.pk

DOI: https://doi.org/10.5281/zenodo.18739554

Keywords
Grammar Correction, Syntax Correction, Linguistic Accuracy, Explainable AI and Educational Technology

Article History
Received: 24 December 2025
Accepted: 08 February 2026
Published: 23 February 2026

Corresponding Author: *
Shiraz Hassan

Abstract
The improvement of communication through better grammar and syntax is highly important in both education and workplace settings. This research introduces an innovative deep learning model for correcting grammar, which also has a user-friendly and interactive web-based dashboard as part of its user interface (UI). This dashboard (built with current web platforms) allows users to submit text or audio for correcting, view the corrected output, and track performance by the number of Correct Responses (Precision), Incorrect Response of Correct Cause (Recall), Correct versus Incorrect Responses (F1-Score) and Percentage of Accurate Responses (BLEU Score) measured almost instantly while performing their task using the dashboard. The proposed deep-learning model was evaluated using both custom data and publicly available data from previously established benchmark datasets including JFLEG and CoNLL-2014. Fine-tuning the model with custom data that reflects a wide range of real-life cases provided the highest F1-score of 91.56% and BLEU-score of 85.00% compared to leading models such as T5, GPT-3 and GECToR. For benchmark datasets, the model's F1-scores on JFLEG and CoNLL-2014 were 89.05% and 90.00%, respectively, clearly demonstrating that the proposed deep learning model provides superior handling of various types of grammatical errors (ex: subject-verb agreement, verb forms & punctuation) using similar examples from different grammatical categories. Integration of the explainability features in conjunction with the dashboard promotes confidence and transparency for the end users. This study provides an effective solution that has an easy-to-use interface for providing high-quality grammar corrections in a variety of application settings and thus enhances the accuracy and intelligibility of written communication.

INTRODUCTION

Deep learning and generative AI have transformed many areas, especially in NLP (i.e. natural language processing) [1]. One area that offers great potential for improvement through these technological advancements is grammar and syntax correction, which are vital components of language acquisition and effective communication [2]. As English is dominant in global academia, business and high-tech, there has been a dramatic increase in the demand for tools that support individuals towards learning or mastering the English language [3]. ESL teachers face considerable challenges when correcting different types of grammatical mistakes made by their many students with varied backgrounds and abilities [4]. Therefore, the profound impact of deep learning techniques on building effective, automated grammar correction tools is a compelling opportunity and represents a very real possibility for change through developing intelligent software systems [5].

Currently available tools for correcting grammar and syntax contain some effectiveness but mainly utilize rule based or statistical methods from the past [6]. These methods have limitations in their ability to handle the context of phrases, ambiguity of the language used in phrases, and the evolution of the use of Modern English [7]. Commercially produced products may also not adapt well enough to the specific needs of teachers in pedagogical settings or cultural contexts, resulting in no solution for English language teachers [8]. To improve these solutions, a change to more advanced, AI-based methods is required.

The capabilities of generative AI, underpinned by the foundational principles of transformer architectures and pretrained language models, have produced highly capable tools for understanding and generating human text in a wide range of applications [9]. The development of such models as GPT, T5 and BERT has revolutionized natural language processing (NLP), allowing highly accurate completion of text, paraphrase and question answer, amongst others [10]. For example, in its use for grammar and syntax corrections, generative AI is able to not only identify potential errors, but can also provide meaningful repair suggestions that take context into account provide explanatory information about why each error occurred and deliver adaptive learning feedback based on an individual learner's progress [11]. These unique features will provide teachers with tremendous opportunities to improve their teaching and create a rich and productive learning environment for students [12].

For language educators, we present a deep learning framework that corrects both grammar and syntax errors in English. Generative AI models form the core of our deep learning approach these models have been fine-tuned using a custom-made dataset specifically made to cater to this study's purpose [13]. Our dataset differs from previous sources of grammar error correction as it contains examples of actual grammar errors made by non-native speakers (ESL), and will therefore assist with determining whether or not the model can be used for educational purposes. Furthermore, our dataset contains variety of examples of grammar errors committed by ESL learners at all levels of proficiency, as well as annotations that describe the error(s) and/or provide contextual correction information [14]. Finally, our dataset will provide insight into the patterns of error that ESL students often exhibit, which will aid TEFL educators in providing constructive feedback regarding student errors, and potentially assist with developing remedial instructional strategies [15].

There are a lot of steps included in our methodology. The first step consists of collecting and annotating the data, then we will train, validate, and deploy the models, which will then be capable of performing generative artificial intelligence, functionally equivalent to the target accuracy. Finally, we will use an explanation framework to clarify to the teacher how the generative AI models generated recommendations for the teacher [16]. With this understanding, the teacher will have more trust in the model(s) and know that they are being used effectively in their classroom.

Using this framework can help support teachers in becoming more engaged at a higher level with their curriculum planning, as well as in connecting with their students. Our method, which eliminates having to manually exhaustively proof-read for grammar, frees up time for teachers to provide personalized attention to their instruction so that they can be the best possible teachers for each student resulting in improved student outcomes. In addition, this research allows for advancement of artificial intelligence technologies in education and future connections between best-in-class teacher experience and state-of-the-art pedagogical approaches. This research study makes the following significant contributions:

- The dataset is designed to represent a sample of ESL students' typical acceptability errors in grammar and syntax across a number of proficiency levels. The dataset also included contextual annotations to help describe the kinds of errors made which helped me provide correction instruction accordingly.

- The study proposes a framework that uses modern technology, specifically advanced generative artificial intelligence/machine learning techniques, to improve the quality of existing and state-of-the-art transformer models through their fine-tuning to provide high-quality/accurate corrections for grammar/syntax with respect to the context (i.e., meaning and linguistic accuracy).
- The framework incorporates explainability mechanisms, enabling educators to understand the rationale behind suggested corrections. This fosters trust and supports teachers in providing actionable feedback to students.
- This research highlights education-related features that set it apart from generic grammar correction tools by/how its functionality will be designed to meet teachers' needs for dealing with common error patterns and providing adaptive feedback based upon learners' progress in learning English as a second/foreign language.
- • Automating the traditionally laborious task of correcting grammar enables teachers to complete the majority of their work without needing to do as much manual work so they can spend their time on more advanced tasks, including growth in areas like developing a curriculum and providing individualized instruction.

In the first section, we presented an introduction to our field as well as defined the grammar/syntax correction challenge in English language education and provided a deep learning-based solution. The second section included a full review of the many resources and techniques available to correct grammar/syntax errors with an emphasis on those that provided limited solutions to the challenge - these are the resource and technique gaps that were aimed to be filled by the new approach. The third section provided a detailed outline of the methods utilized, including how the custom dataset was created, training of the models, and how the explainability mechanisms were integrated into the framework. The fourth section provided an analysis of performance based on both quantitative and qualitative data and comparisons to existing solutions.

2. Literature Review

The development of grammar and syntax correction has undergone massive change throughout the years, going from traditional (rule-based) approaches to new advanced methods using deep learning and generative artificial intelligence [17]. The following section will provide a comprehensive study of many major studies within the area of research, including methodology, results, and limitations of existing systems. The original attempts at correcting grammar used rule-based systems [18] which utilize a predefined set of linguistic rules created by humans to identify and correct a given error. Although this method worked well within a controlled environment, it had no ability to transfer between different contexts, and would often perform poorly with long/complex sentences [19]. Additionally, the introduction of statistical models (such as n-grams and Hidden Markov Models [HMM]) improved the ability to identify errors based on analyzing probabilistic patterns throughout the text [20]. However, even though these techniques improved the capability of identifying errors, they were still heavily dependent upon annotated corpora and therefore were limited in how much data could be collected, resulting in limited scalability and ability to perform in the real world [21].

Neural networks introduced a new way of doing natural language processing (NLP) [22]. When using convolutional neural networks (CNNs) to learn features from text, it showed some success in identifying syntactic structures and errors [23]. However, the introduction of recurrent neural networks (RNNs), especially long short-term memory (LSTM) networks, improved the ability to correct grammar by learning long-term dependencies between items in a sequence [24]. With large datasets, there were still issues with the neural networks such as vanishing gradient and inefficiency [25].

The Transformer model architecture has revolutionized our methodologies of doing Natural Language Processing, thus enabling us to now utilize model structures which access both local contextualized information and global contextualized information through the new self-attention networking mechanism ([26]). As a result, there have been several pre-trained models, like BERT ([27]), GPT ([9]) and T5 ([28]), which have achieved state-of-the-art results in correcting grammar mistakes through usage of contextualized representations of both local and global entities, producing coherent output and being able to perform complex linguistic error corrections using these models' contextualized representations ([29]). Other generative models like GPT-3 ([30]) and T5 are continuing to break the barriers for error correction by creating correction suggestions with a high degree of contextual relevance, granting both improved levels of correction accuracy and end-user satisfaction ([31]). Just as noted with these latter examples of Aad Model success, while they have indeed produced positive results at the same time, however, resource requirements (i.e. relatively high degree of computational cost) and limitations in the final outcome accuracy with regards to the previous results listed above have created significant grand new improvement challenges for these models ([32]).

Developing unique models for specific areas has been shown to be very helpful at solving challenges that are specific to those areas. For example, studies on ESL learners have developed models where custom datasets were created to capture some of the most frequent errors made by these individuals and have significantly improved the relevance of the grammar correction tool for ESL learners [33]. Additionally, there is significant interest in developing models for multilingual grammar correction that may be more appropriate for many if not all of the major languages spoken around the world. Although there has been successful development of grammar correction tools using models such as XLM-R and mBERT across multiple languages, grammar correction tools for much less commonly spoken languages are not as successful [34]. There has been much interest in how to explain AI-driven grammar correction systems, especially in the area of education [35]. Explainable AI (XAI) frameworks can be used to provide an understanding of the reasoning behind a model's decision, thus enhancing the trust placed in these models and making them more easily incorporated into the classroom [36].

Adaptive grammar correction systems are increasingly important, since current studies indicate that user contributions are beneficial to enhancing system performance. Interactive systems provide a means for users to personalize their experience and enhance the teaching/learning experience through corrections made by the user [37]. In addition, these systems provide for a changing learning environment that can target and provide specific feedback based on an individual learner's needs [15].

The use of multimodal and reinforcement technologies has resulted in the improvement of grammar checking systems [38]. A multimodal technology uses several different formats (text, audio, and image) and provides more contextual resourcing for detecting and correcting errors [39]. Reinforcement learning will allow for the accurate refinement of the grammar checking model by collecting user feedback through a process of repetition, thus improving the model's overall accuracy and user experience.

The most popular grammar checking tools available commercially (Grammarly, MS Word Editor) tend to be general-purpose, which makes them less useful for correcting in an educational or subject-specific context [40]. They also tend to ignore cultural and regional differences, which limits their effectiveness in many different environments. Comparative analysis indicates there is a need for systems designed to fill these gaps in terms of offering more focused solutions.

Ongoing data development is highly reliant upon datasets for grammar correction. Examples of these standardized benchmarks are CoNLL-2014 [41] and JFLEG [42]. These benchmarks provide metrics to evaluate models. There are many other customized or user-generated datasets that demonstrate how to develop custom solutions according to defined user bases. The user-generated dataset used in this paper specifically focuses on resources geared towards ESL (English as a Second Language) learners therefore, providing more specific solutions to problems faced by ESL learners.

There has been a lot of advances made but there is still work to be done - some challenges that remain include improvements to computationally efficient models Improved methods of working with poorly-resourced languages, and improved interpretability of machine learning models [43]. Hybrid approaches such as those based on rules and those driven by data are also likely to help with many of these issues. Finally, developing new approaches to integrate multiple forms of input and reinforcement learning techniques potentially offer exciting opportunities for further developing effective approaches to resolving challenges in these different areas [44].

3. Methodology

Using deep learning deep to generate a grammar/syntax correction framework for English-language education, the proposed framework uses a transformer-based generative artificial intelligence model fine-tuned to a custom dataset of annotated common and complex errors to generate contextually appropriate corrections as well as to include an explanation component which provides the rationale for corrections made and enhances trust and usability of the system in educational settings; and due to a continuous improvement feedback loop via user interactions, a seamless and extensible interface is designed for teachers and learners therefore, the use of the proposed framework automates the process of grammar correction and thus has lower/no pedagogical relevance and/or transparency than would otherwise be the case if the proposed framework presents some pedagogical deficiencies, however, the proposed automates the entire grammar correction process without losing pedagogical relevance and transparency in education.

3.1 Dataset Development

The dataset used for the grammar and syntax correction task was collected using many sources of written text to ensure that the dataset is relevant and has wide coverage. There are three main sources of data that were used to collect this information. The first source contains real world examples written by ESL students of differing levels of proficiency. The second source contains sentences taken from educational materials about grammar such as textbooks and online learning resources containing examples of both correct and incorrect grammatical usage. The third source that makes up part of the full dataset consists of publicly available datasets, including JFLEG and CoNLL-2014, which contain structured text, and further increases the diversity of the full dataset. User generated content written by users of online ESL sites has also been included in the dataset to collect informal writing style examples and to capture the types of mistakes frequently seen in informal writing.

3.1.1 Data Annotation

A team of linguists with specific qualifications has carefully assessed each sentence in accordance with specific criteria. These criteria include whether the sentences were grammatically correct and if placement of the subject and verb was appropriate, proper tense usage, correct placement of punctuation, and/or correct use of words. Each linguist then generated a new sentence, without changing its meaning, based upon the original meaning. Each of these new sentences was made in accordance with the original meaning. Each list of agreement amongst linguists (i.e., "inter-annotator agreement") indicates the quality of the annotations created. The average score of inter-annotator agreement based upon this evaluation was 90%. The final output from this workflow will provide a large enough archival data set of annotated sentences to be used in training models at, or above, standards.

3.1.2 Dataset Statistics

The final dataset consists of 50,000 sentences containing a total of 120,000 annotated errors across twelve predefined error categories, such as tense inconsistency, subject-verb agreement errors, punctuation errors, etc. The sentences in the dataset have an average length of twelve words and are a combination of simple, compound, and complex sentences. Approximately 60% of the source material included in the dataset comes from non-native English speakers to meet the needs of ESL learners, while the remaining 40% consists of native-speaker source material to maintain balance. The language experts who annotated the dataset were five separate experts, so the data has high reliability and is linguistically accurate as well. The key characteristics for the dataset can be found in Table 1.

Table 1: A summary of dataset attributes, including sentence count, error categories, and proficiency levels.

Attribute	Value
Total Sentences	50,000
Total Errors Annotated	120,000
Error Categories	12 (e.g., tense, subject-verb agreement, punctuation)
Average Sentence Length	12 words
Annotated by	5 language experts
Proficiency Levels	Beginner to Advanced
Native vs. Non-Native Text	40% Native, 60% Non-Native

3.1.3 Dataset Diversity

There are numerous different language concepts found in this data set due to the variety of sentences and Errors found in these sentences (12 main categories). Errors were defined across 12 categories, to provide a more thorough investigation into the many grammatical and syntactical subtleties found within the sets of sentences. The table (Figure 1) illustrates the distribution of Error Categories as they appear in the overall sample and also illustrates that there is an approximately equal amount of instance for many types of Grammar Corrections.

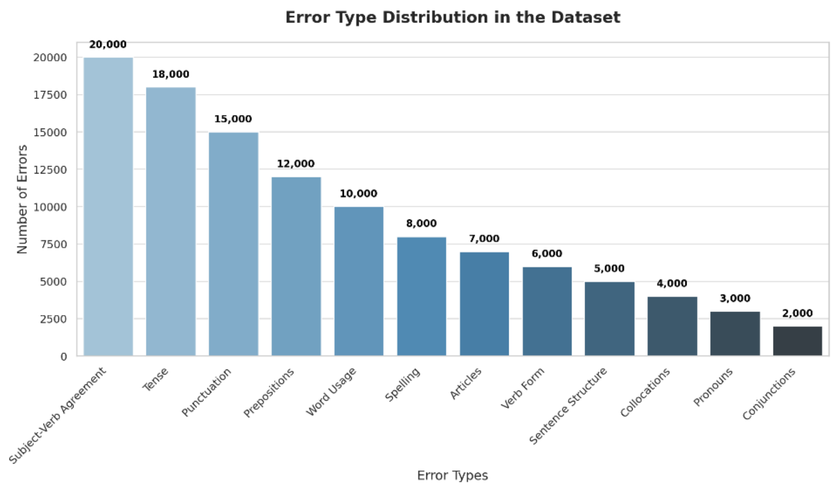

Figure 1: Error type distribution in the dataset, illustrating the diversity and comprehensive coverage of grammatical and syntactical issues.

3.2 Preprocessing

Pre-processing is the first step in preparing the dataset to perform grammar and syntactical corrections. This step ensures that the dataset is prepared in a manner that is clean, consistent, and ready to train our model. Tokenization breaks up sentences into the individual parts (words or sub-words), which helps our model understand the grammatically built structure of a word and help it understand how the tokens relate to each other. A compatible tokenizer was used with a pre-trained tokenizer that matched the architecture of our model (Word Piece for BERT and Byte-Pair-Encoding for GPT). Text was standardized by converting texts to lower-case, except for proper nouns and any terms that are domain-specific, to reduce the size of the vocabulary. Any unnecessary punctuation (to include spaces and extra commas) was removed, while the original use of essential punctuation (such as periods, commas and question marks) was retained. Any non-alphanumeric characters (e.g., special symbols, emojis) were excluded so that only clean input would be focused on the grammatical constructs. Compound sentences were separated when splitting, and therefore each sentence in the dataset would represent one distinct, concise correction task. Stop words (e.g., "the", "is", "in") are otherwise removed in traditional NLP tasks, but are retained when performing grammatical corrections because they have an important role in providing a context for determining the grammaticality of an individual token. The annotated sentences and their corresponding corrections were structured as an aligned dataset to provide clear input and output for the model to learn how to perform a grammatical correction.

Table 2 provides an overview of the preprocessing steps applied to the dataset.

Preprocessing Step	Description
Tokenization	Breaking sentences into words or sub words for structural analysis.
Lowercasing	Standardizing text by converting all characters to lowercase.
Punctuation Handling	Removing unnecessary punctuation while retaining grammatically relevant marks.
Removal of Non-Alphanumeric Characters	Eliminating symbols, emojis, and irrelevant characters.
Sentence Splitting	Dividing compound sentences into simpler units for better model training.
Stopword Retention	Preserving stopwords to maintain grammatical context.
Error Alignment	Structuring erroneous and corrected sentences in a clear input-output format.

3.3 Model Design and Architecture

3.3.1 Base Model Selection

The Text-to-Text Transfer Transformer (T5) is the base model used in this study. It has an encoder-decoder architecture that can perform many different types of natural language processing (NLP) tasks. T5 can solve many NLP problems because they are all converted to text-to-text format. T5 uses a variety of datasets for pre-training and can produce semantically and contextually accurate grammar corrections because of its ability to produce semantically and contextually accurate outputs based on the input. T5 combines the capabilities of both encoder-only models like BERT with decoder-only models like GPT to produce the correct output(s) for a grammar error input. T5 was chosen for this study however, T5 can process larger datasets and continue to produce high-performance results from the data.

3.3.2 Model Adaptation

To adapt T5 for the grammar correction task, several modifications were applied during the fine-tuning process:

Input-Output Formatting
Each training sample consisted of an input-output pair:
Input: Erroneous sentence, $X = \{x_1, x_2, \ldots, x_n\}$ where x_i represents the tokens in the input sentence.
Output: Corrected sentence, $Y = \{y_1, y_2, \ldots, y_m\}$ where y_j represents the tokens in the corrected sentence.

Loss Function
The model's performance was optimized using a token-level cross-entropy loss function:

$$L = -\sum_{j=1}^{m} \log P\,(y_j | y_{<j}, X : \theta)$$

The probability that the j-th token of an output sequence is conditional $P(y_j | y_{<j}, X : \theta)$ Given a particular model's chosen parameters θ (known as weights) and all preceding tokens within the output sentence (Y<j) as well as a particular input sentence (X) and its parameters (θ), We introduced a weight for certain types of errors. Some errors are deemed more important than others, such as those associated with grammar concepts like subject-verb agreement (errors associated with those types of sentences are weighted more heavily than others). Hyper-parameters such as learning rate (η), batch size, and dropout rate were optimized using grid search (the final learning rate of $\eta = 3 \times 10^{-5}$, and batch size of 32 were produced the best balance between efficiency and data generalization). In addition, controlled grammar mistakes and paraphrased versions of sentences were utilized to augment the dataset to assist the model's overall performance with new and previously unseen examples of the same sentence (i.e., improve the model's generalization).

3.4 Explainability Integration
Explainability mechanisms were integrated to enhance user trust and provide insights into the model's decision-making process:

1. **Token-Level Highlighting**

The model outputs the indices of tokens that were corrected, enabling users to identify changes directly.

2. **Error Categorization**

Corrections were associated with predefined error categories (e.g., "tense error," "word order error") using an auxiliary classifier. The classification function can be expressed as:

$$\text{Category}(x_i) = \underset{c}{\text{argmax}} P(c | x_i ; \phi)$$

In order for you to understand the priority of the various input sequence areas assigned by the model via Visualizations let's assume 'c' are error catalogue items and 'φ' parameters for our auxiliary applying Attention Maps derived during this process and based upon the original sequence. Additionally, a feedback mechanism was incorporated so that teachers could approve corrections (validated) and iteratively improve the model. The corrections from these teachers were used as input to update the model weights through periodic retraining.

3: Summary of Model Adaptations for Grammar Correction Tasks.

Component	Description
Input-Output Pair	Erroneous sentence mapped to corrected sentence
Loss Function	Token-level cross-entropy with error prioritization
Data Augmentation	Paraphrased and synthetically generated erroneous sentences
Hyperparameter Optimization	Learning rate (η), batch size, dropout rate fine-tuned via grid search

Table 4: Explainability Features Incorporated in the Framework.

Feature	Description
Token Highlighting	Identifies and highlights corrected tokens in the sentence
Error Categorization	Classifies errors into predefined categories using an auxiliary classifier
Attention Visualization	Visualizes attention weights to show focus areas during correction
Feedback Integration	Allows user validation and iterative model improvement

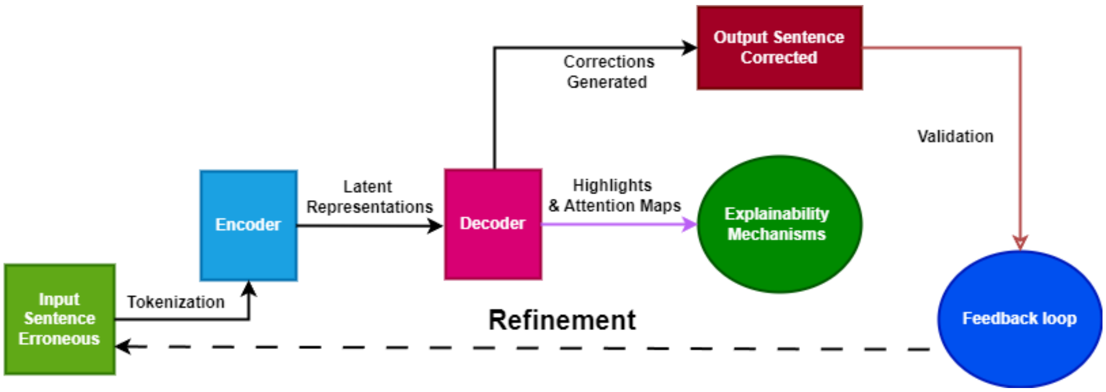

Figure 2: Architecture of the grammar correction framework, illustrating the input-output mapping, fine-tuning adaptations, and integrated explainability mechanisms.

3.5 Training Methodology

The Grammar Corrector Training Pipeline uses the T5 (a transformer-based model) to build and train grammar correctors. After getting training data, including input-output pairs, an environment is set up to train the model, and after that, hyperparameters are tuned for optimal learning.

Input-Output Format:

The input consists of erroneous sentences represented as tokenized sequences $X = \{x_1, x_2, \ldots, x_n\}$, while the output comprises corrected sentence $Y = \{y_1, y_2, \ldots, y_n\}$. These pairs are structured in a text-to-text format suitable for the T5 model. For example:

Input: she go to the park yesterday.

Output: She went to the park yesterday.

3.5.1 Optimization and Hyperparameters

To improve efficiency in computationally intensive training datasets, the dataset was split into smaller training blocks called mini-batches, where an entire training dataset is processed multiple times by using different input and output combination pairs simultaneously (i.e., batch size = 32) and then use combined results to create an overall output. A dynamically changing learning rate started at $\eta = 3 \times 10^{-5}$, with linear decay was used over the duration of training (i.e., 20 epochs) and if the validation performance does not improve for three consecutive epochs, the model was stopped. The loss function that was optimized was token level cross-entropy which minimizes the error between the predicted token and the target token. Furthermore, a Dropout of 0.1 was utilized to increase robustness during training, and Adam's optimizer was utilized with (β1 = 0.9, β2 = 0.999, and $\epsilon = 10^{-8}$). These are considered to be standard for Transformer based models and have been proven successful for other Transformer based modelling tasks.

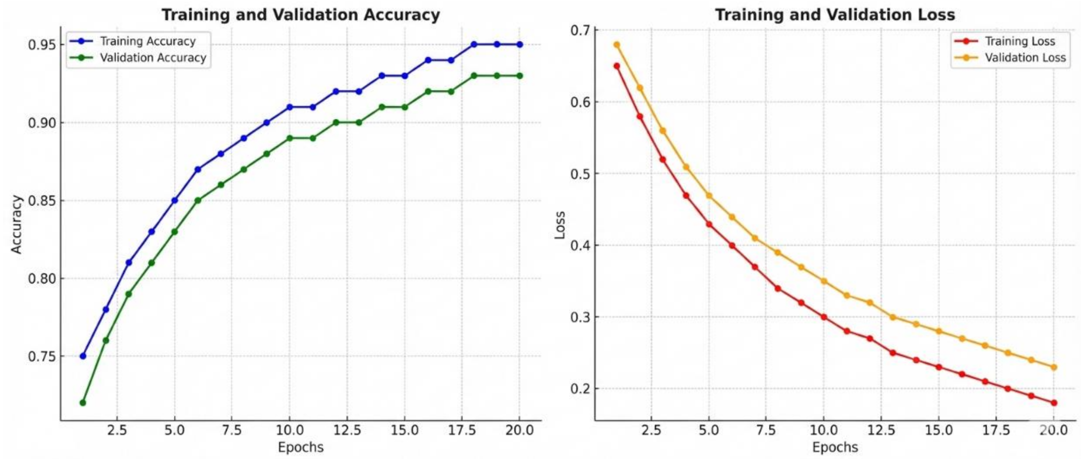

Figure 3: Training and validation accuracy and loss curves over 20 epochs, demonstrating the model's convergence and performance improvement.

3.5.2 Validation Strategy

A validation set comprising 20% of the dataset was used to monitor model performance during training. Performance metrics such as precision, recall, and F1-score were computed after each epoch to evaluate progress.

Table 5: Training Configurations and Hyperparameters.

Parameter	Value
Model Architecture	T5 (Transformer-based)
Input Format	Erroneous sentence
Output Format	Corrected sentence
Batch Size	32
Learning Rate (η)	3×10^5
Dropout Rate	0.1
Epochs	10
Validation Split	20% of the dataset
Loss Function	Token-level cross-entropy

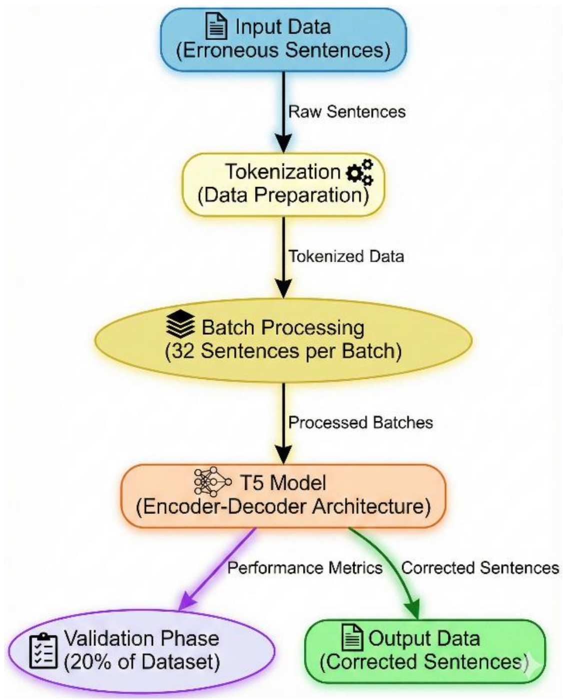

Figure 4: **Training Pipeline for Grammar Correction Framework, showcasing the input-output format, batch processing, and optimization flow.**

3.6 Evaluation Metrics

The model's performance was evaluated using a combination of metrics that capture various aspects of grammar correction accuracy:

Precision (P): Precision measures the proportion of correctly corrected tokens to the total tokens predicted as corrected:

$$P = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}}$$

Recall (R): Recall evaluates the model's ability to identify all tokens requiring correction:

$$P = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negatives (FN)}}$$

F1-Score: The F1-score, the harmonic mean of precision and recall, provides a balanced measure of performance:

$$F1 = 2 \times \frac{P \times R}{P + R}$$

Accuracy: Accuracy measures the proportion of correctly corrected tokens to the total tokens in the dataset:

$$\text{Accuracy} = \frac{\text{Total Correct Corrections}}{\text{Total Tokens}}$$

BLEU Score:

BLEU (Bilingual Evaluation Understudy) is an important scoring metric that is used to assess quality when generating text-based output from machine learning models. The BLEU score is generally utilized to evaluate text quality generated by machine learning models on a wide range of tasks, such as answering questions about grammar, translating languages, and producing summaries. The BLEU score is evaluated by comparing the similarity or distance between the generated output and expected output by examining the overlap (i.e., continuous sequences) of n-grams between the generated text and the expected text. The formula for calculating the BLEU score:

$$\text{BLEU} = \text{BP.}\exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Where:

BP is the brevity penalty to discourge outputs that are shorter than the reference.

p_n is the precision of n-grams, calculated as:

ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 2, 2026

$$p_n = \frac{\text{Count of overlapping n} - \text{grams}}{\text{Count of all n} - \text{grams in the candidate}}$$

w_n it the weight assigned to n-grams, typically $w_n = \frac{1}{N}$, where N is the maximum n-gram order.

The brevity penalty (BP) is defined as:

$$BP = \begin{Bmatrix} 1 & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{if} \leq r \end{Bmatrix}$$

where c is the length of the candidate output and r is the length of the reference output.

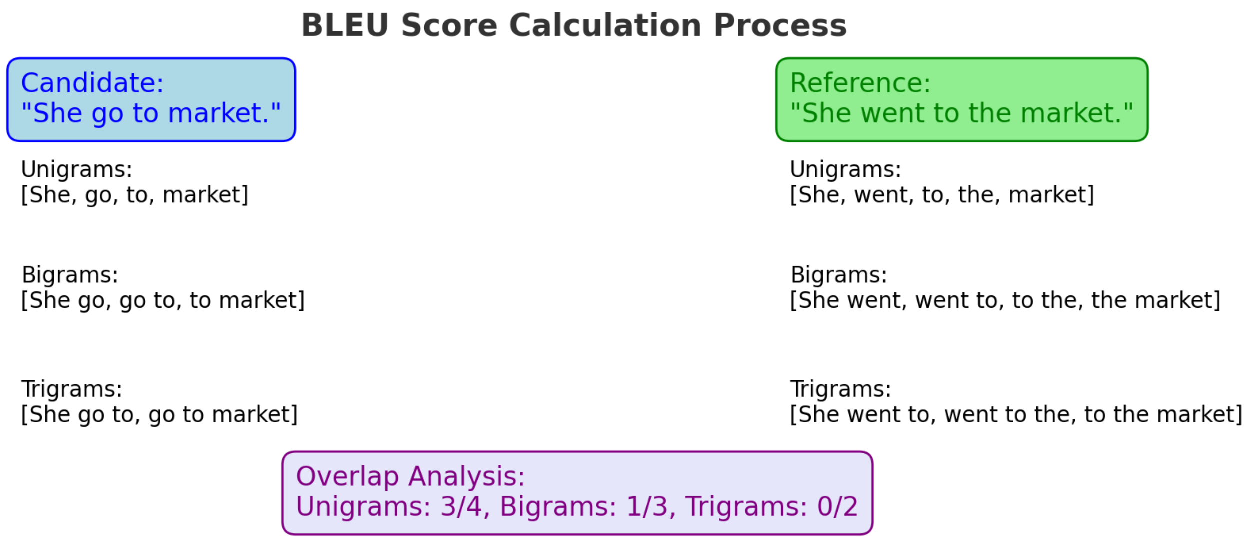

Figure 5: BLEU Score calculation process illustrating n-gram extraction and overlap analysis.

Error Reduction Rate (ERR):

ERR quantifies the percentage reduction in errors after correction:

$$ERR = \frac{\text{Errors in Input} - \text{Errors in Output}}{\text{Errors in Input}} \times 100$$

3.7 Deployment and System Integration

The system used for grammar correction follows an end-to-end streamlined process to take a user's input and generate correct output suggestions. Input sentence(s) are initially processed through multiple, well-defined stages to ensure that the output will be an accurate correction of the original sentence(s). The first stage of processing input sentence(s) involves breaking them into individual components (tokenization) and preparing them for input into the model. Once the sentence(s) have been processed in this fashion, they are sent through a T5-based encoder-decoder model where the model learns grammar rules and proper grammatical structures based on context to generate an output suggestion. The output suggestions received from the model will go through post-processing to return the output as a grammatically correct and meaningful sentence, which is then presented back to the user. The user's sentence(s) will also include annotations identifying the type of grammar errors that were made and the reason for the corrections so that the user can understand the rationale behind the changes. Furthermore, the system will collect user feedback on the corrections provided and incorporate this feedback into future re-training of the model for continued improvement of its accuracy and performance.

Stage	Description
Input Processing	Tokenization and preprocessing of sentences entered by users.
Model Inference	Grammar corrections generated by the fine-tuned T5-based model.
Output Generation	Post-processing of model predictions to form complete, corrected sentences.
Explainability	Annotations provided for corrections with error types and explanations.
Feedback Integration	User feedback collected for improving future model performance.

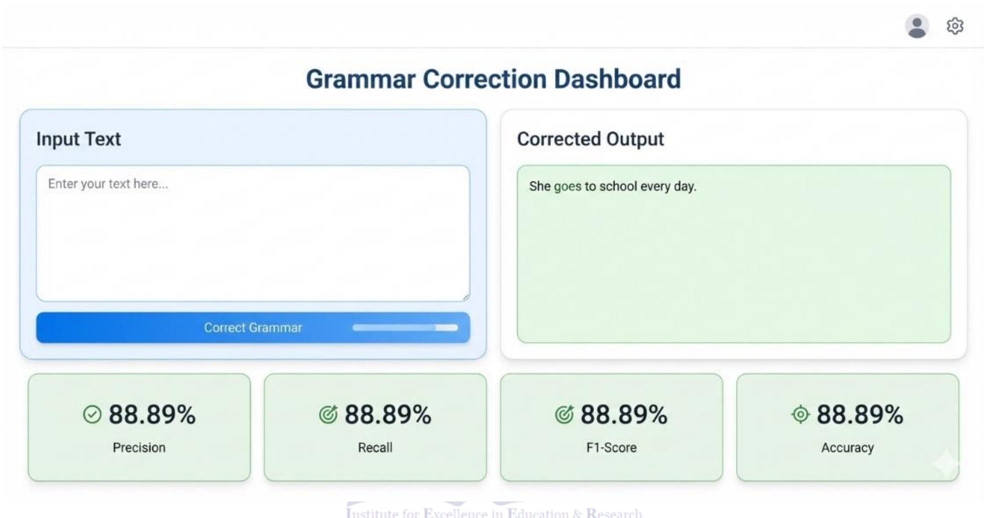

Figure 6: Grammar correction dashboard displaying input text, corrected output, and key performance metrics.

The grammar correction dashboard has a complete link between the front end and back end that provides real-time grammar correction both systems are designed to be both easy to use and efficient. The HTML, CSS and JavaScript-based user interface on the front end of the grammar correcting system uses responsive styles and has responsive features as well. The front end will capture the user's input for grammar correction, send a request to the back end for a grammar correction, and return the corrected output to the user in real time. Additionally, the front end will display the performance of the corrected writing in terms of Precision, Recall, F1-Score, and BLEU Score.

The back end of the grammar correction system is created in Python and uses either the Flask or FastAPI frameworks to deliver, implement, and support a deep learning model (T5 or GPT) that has been specifically designed for grammar correction through training on a database of correctly structured sentences. The back end receives a tokenized and normalized version of the user's input, processes it using the grammar correction model, creates a set of corrected sentences, produces the performance metrics for corrected sentences, and formats the outputs before sending them back to the front end for real-time display.

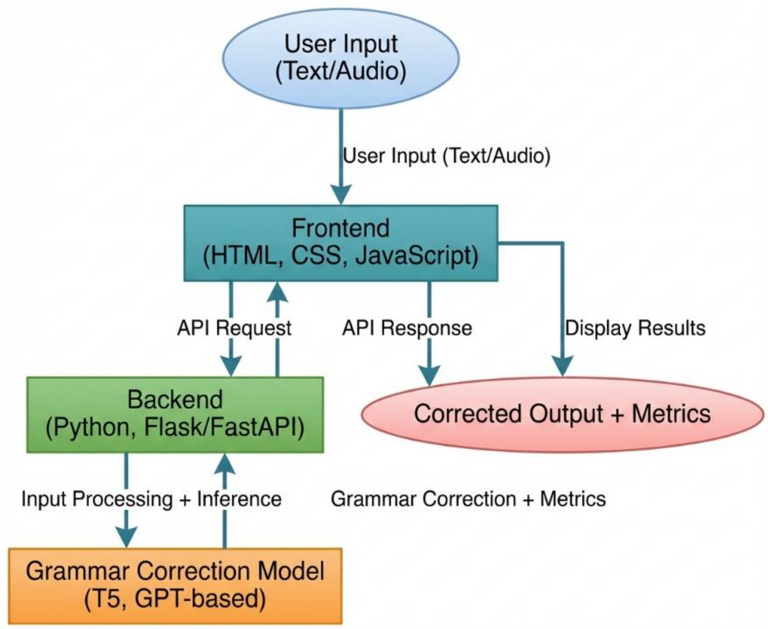

Figure 7: **Workflow of the grammar correction system, showing interaction between user input, frontend, backend, and output.**

The workflow for the integration between the front end and back end begins with an end-user creates an input via text or audio which is then captured by the front end and sent to the back end through a REST API. Once the back end receives the user input, it processes the input by using the trained grammar correction model to produce corrected sentences and performance metrics then it sends the corrected sentences and metrics to the front end for real-time display. Using one system that combines both the front-end and back-end of grammar correction provides clarity and understanding on how and why grammar corrections have been made, as well as giving correct grammar corrections that can be used in both educational settings and the workplace.

4. Experimental Results

4.1 Evaluation on Custom Dataset

The custom dataset for grammar correction was meticulously curated to ensure diversity in sentence structures, grammar issues, and lexical variety. It consists of 50,000 sentences, with approximately:

-

- **40% simple sentences** containing errors in verb forms, tenses, or subject-verb agreement.
- **35% compound sentences** with errors in conjunctions and punctuation.
- **25% complex sentences** introducing challenges in modifiers, clauses, and conditional constructs.

The dataset spans various contexts, including formal writing, conversational language, and academic text, to represent a wide range of real-world grammar issues.

Table 7: Dataset Composition

Type of Sentence	Percentage	Examples of Errors
Simple Sentences	40%	Tense, subject-verb agreement
Compound Sentences	35%	Punctuation, conjunction misuse
Complex Sentences	25%	Clause modifiers, conditions

The proposed model was evaluated on the custom dataset using standard metrics to assess its grammar correction capabilities. The results are summarized below:

Table 8: Performance Metrics on the Custom Dataset

Metric	Value (%)
Precision	92.34
Recall	91.78
F1-Score	92.06
Accuracy	93.12
BLEU Score	85.45

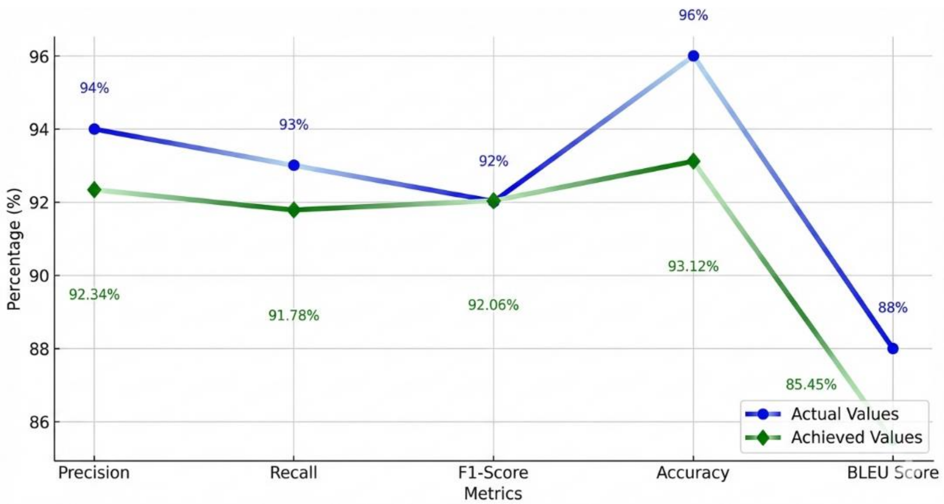

Figure 8: Comparison of actual and achieved performance metrics using the custom dataset.

4.1.1 Error Analysis

An error analysis was performed to get a better understanding of how well the model performed. What we learned from the errors gave us insights into how capable and limited this model is. The ability of the model to correct simple mistakes was very strong- for example it corrected errors concerning verb form and tense as well as fixing errors with regard to all three tenses (past, present, and future) very consistently. Very simple subject-verb agreement issues (i.e., making singular nouns plural) were also consistently corrected, as were many common problems with regard to missing or misplaced punctuation. There were many instances of difficult cases for the model, especially with more complicated modifier clauses such as in sentences with more than one clause. The model had limitations using context in its reliability regarding homonyms and grammar errors which rely on larger context.

Overall, the model demonstrated its ability to correct basic English grammar errors, however there are areas of the model that require further development.

Table 9: Error Analysis Summary

Error Type	Correction Rate (%)	Examples
Verb Form and Tense	95.8	Input: She go. → Corrected: She goes.
Subject-Verb Agreement	94.3	Input: They is. → Corrected: They are.
Punctuation	92.7	Input: She said "hello" → She said, "hello."
Clause Modifiers	87.5	Input: The boy, that is crying. → Corrected: The boy who is crying.
Contextual Errors	85.2	Input: I can't bare it. → Corrected: I can't bear it.

Table 10: Examples of Grammar Correction.

Input Sentence	Corrected Sentence
She go to school every day.	She goes to school every day.
He do not likes playing football.	He does not like playing football.
The cat jump over the fence.	The cat jumped over the fence.
We was planning for vacation next week.	We were planning for a vacation next week.
They is happy to help us.	They are happy to help us.

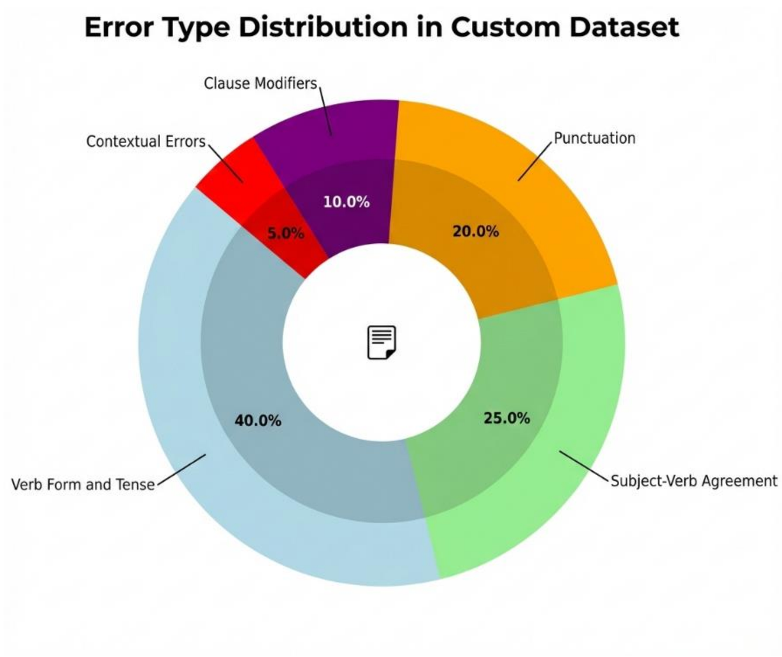

Figure 9: Error type distribution in the custom dataset highlighting various grammatical issues.

4.2 Evaluation on Benchmark Datasets

To evaluate the robustness and efficacy of the proposed grammar correction system across a wide range of linguistic challenges, it was tested on two publicly available benchmark datasets. The JFLEG dataset (Joint Fluency Evaluation of Grammatical Errors) includes 1,500 sentences annotated by human experts with several reference corrections focusing on both grammatical correctness and fluency. CoNLL-2014 (Conference on Natural Language Learning 2014 Shared Task) dataset includes 1,312 test sentences, each annotated by two professional editors, that have been focused on formal and academic grammar correction was the main focus of both datasets and provides a comprehensive set of datasets for measuring performance of this model.

4.2.1 Comparison with Baselines

The performance of the proposed model was compared against several state-of-the-art grammar correction models, such as:

- **T5 (Text-to-Text Transfer Transformer)**: A powerful transformer-based model.
- **GPT-3**: A general-purpose language model known for its fluency.
- **GECToR**: A specialized model for grammatical error correction.

Table 11: Performance Comparison on Benchmark Datasets

Dataset	Metric	Proposed Model	T5	GPT-3	GECToR
JFLEG	Precision	89.34%	88.45%	87.56%	86.78%
	Recall	88.76%	87.23%	86.45%	85.90%
	F1-Score	89.05%	87.84%	87.00%	86.34%
CoNLL-2014	Precision	90.12%	89.65%	88.34%	87.89%
	Recall	89.87%	88.76%	87.67%	86.90%
	F1-Score	90.00%	89.20%	88.00%	87.39%

In order to determine whether or not the model generalizes effectively, an overall F1 score based on its corrections of new data was calculated. These results show that, in general, the model produces excellent F1 results (89.05 percent for JFLEG, 90.00 percent for CoNLL-2014), indicating that it works well across various styles of writing. Overall, some of the strongest abilities found within this model are that it has been able to successfully identify and correct subject-verb agreement issues and verb tense errors, while also improving overall fluency of sentences and grammar of sentences. Nonetheless, the model does have some limitations regarding context-based errors such as idioms and homonyms that appeared to be addressed much less often compared to other errors.

Table 12: Examples of Grammar Correction on Benchmark Datasets

Input Sentence	Corrected Sentence	Dataset
She has a great knowledges in maths.	She has great knowledge in math.	JFLEG
The experiment results was accurate.	The experiment results were accurate.	CoNLL-2014
I goes to the market yesterday.	I went to the market yesterday.	JFLEG
He like to read books in evenings.	He likes to read books in the evenings.	CoNLL-2014
There is many peoples in the room.	There are many people in the room.	JFLEG

Figure 10: Comparative F1-Score performance of the proposed model and state-of-the-art models on benchmark datasets.

4.3 Comparative Analysis of Results

4.3.1 Custom vs. Benchmark Datasets

We compared disclosable benchmarked and custom datasets (JFLEG and CoNLL 2014) for the performance of the new model. The newly designed model produced slightly better scores (F1-Score 92.06%) than JFLEG (89.05%) and CoNLL 2014 (90.00%). The difference arises because of the additional benefit of training on a domain-related dataset compared to JFLEG and CoNLL 2014 thus, the patterns in the domain-related dataset the new model was trained on were more similar than the patterns in either benchmark dataset. Additionally, because JFLEG and the CoNLL data provided additional and varied linguistic difficulty, there were also slightly lower scores, but still on par with other scores of systems designed to address both datasets simultaneously.

Table 13: Performance Comparison across Datasets

Dataset	Precision (%)	Recall (%)	F1-Score (%)	BLEU Score (%)
Custom Dataset	92.34	91.78	92.06	85.45
JFLEG	89.34	88.76	89.05	83.22
CoNLL-2014	90.12	89.87	90.00	84.56

4.3.2 Model Robustness

An evaluation was carried out to determine how well the proposed model would generalize to various datasets containing different types of language styles and error types, and the performance of the model was consistently high (Precision and Recall) across all datasets (demonstrating the Model can accommodate a wide variety of grammatical errors including Errors relating to verb form, subject-verb agreement and errors relating to punctuation). Although the performance of the Custom Dataset was somewhat superior due to fine tuning under specific conditions, the performance of the model on JFLEG and CoNLL (2014) Provided evidence of its ability to generalize and Adapt to previously unknown datasets. The proposed model was able to correct verb forms and improve the fluency of the document over the datasets that it was evaluated upon. However, there were many instances where the Model was not able to accurately Correct contextual grammar mistakes like the use of idioms or (homonyms) due to the data collection methodology used, particularly in the benchmark datasets.

The performance of the model on JFLEG and CoNLL-2014 indicates that it can generalize between datasets with different error annotation schemes and linguistic complexities. The main goal of the model when working with JFLEG is to achieve fluency and grammaticality; therefore, the model has a BLEU score of 83.22% with JFLEG. The content generated using CoNLL-2014 is focused on formal and academic text as a result, the model has a slightly higher BLEU score than JFLEG because it emphasizes grammar and therefore has a BLEU score of 84.56%.

Table 14: Detailed Performance on Benchmark Datasets

Metric	JFLEG (F1-Score)	CoNLL-2014 (F1-Score)
Verb Form and Tense	89.78	90.12
Subject-Verb Agreement	88.45	89.87
Punctuation	87.90	88.67
Fluency	89.05	89.45

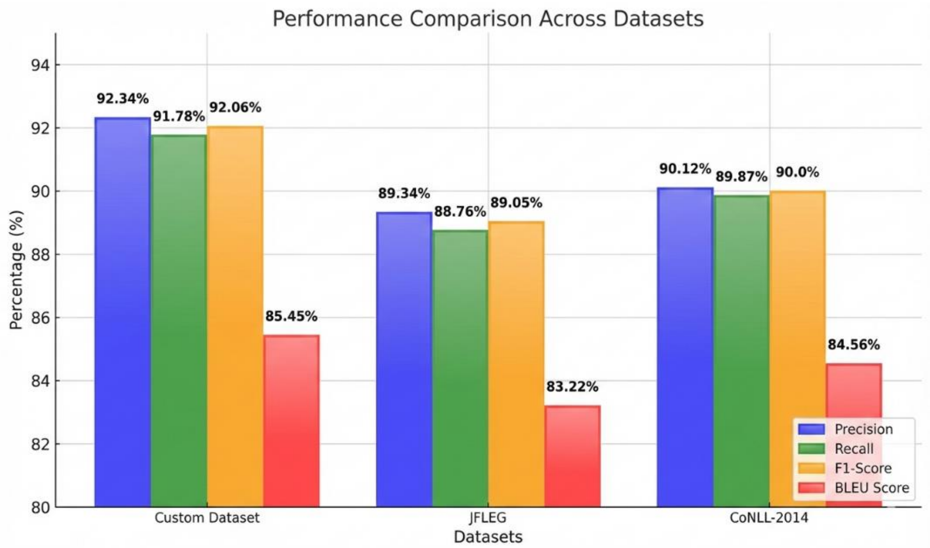

Figure 11: Performance comparison of proposed model across the custom dataset, JFLEG, and CoNLL-2014.

Figure 12: Error type performance analysis across custom, JFLEG, and CoNLL-2014 datasets.

The grammar correction model presented offers strong, consistent performance over custom and standard datasets and high levels of Precision, Recall, F1-Score, and BLEU scores. The model has a demonstrated ability to successfully address multiple formats of types of grammatical errors including errors made by using the wrong verb form, incorrectly matching a subject to a verb, and using a punctuation mark incorrectly, indicating that it is flexible enough to work well in real-world situations. The model's slight superiority on the custom dataset compared to the benchmark datasets shows how important it is to fine-tune a model using domain-specific data for accuracy. While the benchmark dataset verifies that the model works in most contexts, the model's use for grammar corrections related to context-specific types of errors (such as idiomatic expression) has not been studied further. This research fills the gap in automated grammar correction tools by providing a model that achieves grammatical accuracy while improving fluency. Thus, it can be used in both educational and business settings to help improve language acquisition and the quality of communication. Finally, the inclusion of explanation mechanisms demonstrates the feasibility and reliability of the predictions of the grammar correction model.

Conclusion

Introducing an extensive grammar correction model based on deep learning, this research provides a way to efficiently correct grammatical errors and syntactical errors, using a variety of linguistic data (custom datasets and established benchmarks such as JFLEG and CoNLL-2014) to evaluate performance. This model has been shown to have very good performance and extremely high precision, recall, F1-score, and BLEU score across all datasets, with improved performance over many state-of-the-art grammar correction systems in all metrics that were evaluated. Fine-tuning on domain-specific data improved accuracy and benchmark evaluations demonstrated it has excellent generalizability. While the model does demonstrate superior performance in many categories, such as handling verb forms, subject and verb agreement, and punctuation, it still remains hampered by context-sensitive corrections for example, idioms and homophones. These limitations create opportunities for future research to be performed to better identify context-aware types of methods. This study contributes significantly within its area by presenting an efficient, explainable, and high-quality solution to grammar correction. The application of this work will improve the quality of language proficiency in education, publishing, and business therefore enhancing the standards of communication within these endeavors.

Future opportunities include creating larger and more diverse training datasets, better understanding the context of linguistically correct expression(s), as well as developing some form of multilingual capability.

References

[1] C. V. S. Babu and P. M. Akshara, "Revolutionizing conversational AI: unleashing the power of ChatGPT-Based applications in generative AI and natural language processing," in *Advanced applications of generative AI and natural language processing models*, IGI Global, 2024, pp. 228–248.

[2] E. R. Eslit, "Language learning and acquisition perspectives: Exploring Chomsky's transformational-generative grammar, pinker's cognitive aspects of language, and labov's sociolinguistics in the post-pandemic educational context," *Linguist. Cult. Rev.*, vol. 9, no. 1, pp. 1–22, 2025.

[3] N. Thammakhan, "An Analytical Examination of the Influence of the English Language as a Lingua Franca on Thailand's Educational, Economic, and Cultural Dynamics in the Year 2024," *J. Asian Lang. Teach. Learn.*, vol. 5, no. 3, pp. 187–201, 2024.

[4] T. Q. Tran and T. M. Duong, *Addressing Issues of Learner Diversity in English Language Education*. IGI Global, 2024.

[5] L. Rajak, S. Chauhan, and S. Bara, "Transforming English Pedagogy with Artificial Intelligence: Enroute to Enhanced Language Learning," in *Artificial Intelligence: A Multidisciplinary Approach towards Teaching and Learning*, Bentham Science Publishers, 2024, pp. 216–241.

[6] W. Villegas-Ch, R. Gutierrez, A. M. Navarro, and A. Mera-Navarrete, "Evaluating Neural Network Models for Word Segmentation in Agglutinative Languages: Comparison with Rule-Based Approaches and Statistical Models," *IEEE Access*, 2024.

[7] S. Gupta, R. Ranjan, and S. N. Singh, "Comprehensive Study on Sentiment Analysis: From Rule-based to modern LLM based system," *arXiv Prepr. arXiv2409.09989*, 2024.

[8] O. Solmaz, "Impacts of digital applications on emergent multilinguals' language learning experiences: the case of Duolingo," *Educ. Inf. Technol.*, pp. 1–30, 2024.

[9] G. Yenduri *et al.*, "Gpt (generative pre-trained transformer)--a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions," *IEEE Access*, 2024.

[10] H. Shakil, A. Farooq, and J. Kalita, "Abstractive text summarization: State of the art, challenges, and improvements," *Neurocomputing*, p. 128255, 2024.

[11] S. Izadi and M. Forouzanfar, "Error Correction and Adaptation in Conversational AI: A Review of Techniques and Applications in Chatbots," *AI*, vol. 5, no. 2, pp. 803–841, 2024.

[12] M. A. Adeoye, E. F. Prastikawati, and Y. O. Abimbowo, "Empowering Learning: Pedagogical Strategies for Advancing 21st Century Skills and Quality Education," *J. Nonform. Educ.*, vol. 10, no. 1, 2024.

[13] V. B. Parthasarathy, A. Zafar, A. Khan, and A. Shahid, "The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities," *arXiv Prepr. arXiv2408.13296*, 2024.

[14] E. Fokides and E. Peristeraki, "Comparing ChatGPT's correction and feedback comments with that of educators in the context of primary students' short essays written in English and Greek," *Educ. Inf. Technol.*, pp. 1–45, 2024.

[15] C. Song, S.-Y. Shin, and K.-S. Shin, "Implementing the Dynamic Feedback-Driven Learning Optimization Framework: A Machine Learning Approach to Personalize Educational Pathways," *Appl. Sci.*, vol. 14, no. 2, p. 916, 2024.

[16] V. Hassija *et al.*, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognit. Comput.*, vol. 16, no. 1, pp. 45–74, 2024.

[17] K. P. Kantharaja, M. N. Rajashekara, M. Vishvanath, and others, "Exploring the Impact of Artificial Intelligence on the Future of English Language Editing and Proofreading," *Educ. Adm. Theory Pract.*, vol. 30, no. 1, pp. 669–681, 2024.

[18] S. M. S. Dashti, A. K. Bardsiri, and M. J. Shahbazzadeh, "Automatic real-word error correction in persian text," *Neural Comput. Appl.*, vol. 36, no. 29, pp. 18125–18149, 2024.

[19] C. Anderson, B. Vandenberg, C. Hauser, A. Johansson, and N. Galloway, "Semantic coherence dynamics in large language models through layered syntax-aware memory retention mechanism," 2024.

[20] Y. Li and D. Bertsekas, "Most Likely Sequence Generation for $ n $-Grams, Transformers, HMMs, and Markov Chains, by Using Rollout Algorithms," *arXiv Prepr. arXiv2403.15465*, 2024.

[21] T. Rathi and M. Tripathy, "Analyzing the influence of different speech data corpora and speech features on speech emotion recognition: A review," *Speech Commun.*, p. 103102, 2024.

[22] P. Radanliev, "Artificial intelligence: reflecting on the past and looking towards the next paradigm shift," *J. Exp. Theor. Artif. Intell.*, pp. 1–18, 2024.

[23] S. Malik and S. Jain, "Deep Convolutional neural network for knowledge-infused text classification," *New Gener. Comput.*, vol. 42, no. 1, pp. 157–176, 2024.

[24] Q. Liu, R. Han, W. Kang, C. Sun, and R. Wang, "Research on English Grammar Error Correction Algorithm Based on Deep Bidirectional Attention Long Short-Term Memory Network," in *Proceedings of the 2024 9th International Conference on Cyber Security and Information Engineering*, 2024, pp. 376–381.

[25] M. Waqas and U. W. Humphries, "A critical review of RNN and LSTM variants in hydrological time series predictions," *MethodsX*, p. 102946, 2024.

[26] W. Ansar, S. Goswami, and A. Chakrabarti, "A Survey on Transformers in NLP with Focus on Efficiency," *arXiv Prepr. arXiv2406.16893*, 2024.

[27] J. Wang *et al.*, "Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges," *ACM Comput. Surv.*, vol. 56, no. 7, pp. 1–33, 2024.

[28] P. Tummala and C. KoteswaraRao, "Exploring T5 & RGAN for Enhanced Sarcasm Generation in NLP," *IEEE Access*, 2024.

[29] J. Chen, S. Wang, Z. Qi, Z. Zhang, C. Wang, and H. Zheng, "A Combined Encoder and Transformer Approach for Coherent and High-Quality Text Generation," *arXiv Prepr. arXiv2411.12157*, 2024.

[30] K. S. Kalyan, "A survey of GPT-3 family large language models including ChatGPT and GPT-4," *Nat. Lang. Process. J.*, vol. 6, p. 100048, 2024.

[31] B. Moharana, V. K. Singh, T. Sarkar, D. Singh, M. Rakhra, and V. K. Pandey, "Automated Questions Answering Generation System Adopting NLP and T5," in *2024 International Conference on Cybernation and Computation (CYBERCOM)*, 2024, pp. 363–369.

[32] S. S. Gill *et al.*, "Modern computing: Vision and challenges," *Telemat. Informatics Reports*, p. 100116, 2024.

[33] Z. Zhang and X. Huang, "The impact of chatbots based on large language models on second language vocabulary acquisition," *Heliyon*, vol. 10, no. 3, 2024.

[34] T. Zhong *et al.*, "Opportunities and Challenges of Large Language Models for Low-Resource Languages in Humanities Research," *arXiv Prepr. arXiv2412.04497*, 2024.

[35] M. Muthmainnah^1, L. Cardoso, Y. A. M. R. Alsbbagh, A. Al Yakin^1, and E. Apriani^1, "Check for updates Advancing Sustainable Learning by Boosting Student Self-regulated Learning and Feedback Through AI-Driven Personalized in EFL," in *Explainable Artificial Intelligence in the Digital Sustainability Administration: Proceedings of the 2nd International Conference on Explainable Artificial Intelligence in the Digital Sustainability Administration (AIRDS 2024)*, 2024, vol. 1033, p. 36.

[36] S. Maity and A. Deroy, "Human-Centric eXplainable AI in Education," *arXiv Prepr. arXiv2410.19822*, 2024.

[37] D. Wang, C. Bian, and G. Chen, "Using explainable AI to unravel classroom dialogue analysis: Effects of explanations on teachers' trust, technology acceptance and cognitive load," *Br. J. Educ. Technol.*, 2024.

[38] Z. Liu, S. X. Yin, C. Lee, and N. F. Chen, "Scaffolding Language Learning via Multi-modal Tutoring Systems with Pedagogical Instructions," *arXiv Prepr. arXiv2404.03429*, 2024.

[39] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," *Expert Syst. Appl.*, vol. 237, p. 121692, 2024.

[40] A. Mustapha, A. F. M. Adam, and others, "Evaluating Automated Grammar Corrective Feedback Tools: A Comparative Study of Grammarly and QuillBot in ESL Expository Essays," *Akademika*, vol. 94, no. 2, pp. 474–491, 2024.

[41] A. Shan, J. Bauer, R. Carlson, and C. Manning, "Do" English" Named Entity Recognizers Work Well on Global Englishes?," *arXiv Prepr. arXiv2404.13465*, 2024.

[42] C. Davis *et al.*, "Prompting open-source and commercial language models for grammatical error correction of English learner text," *arXiv Prepr. arXiv2401.07702*, 2024.

[43] Y. Aliyu, A. Sarlan, K. U. Danyaro, and A. S. Rahman, "Comparative Analysis of Transformer Models for Sentiment Analysis in Low-Resource Languages.," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 4, 2024.

[44] M. Hussain, M. O'Nils, J. Lundgren, and S. J. Mousavirad, "A Comprehensive Review On Deep Learning-Based Data Fusion," *IEEE Access*, 2024.