Innovative Informatics and Artificial Intelligence Research (IIAIR)
Vol. 1, Issue 1, pp. 26-34, 2025
Received 21 Jan 2025; accepted 10 Apr 2025; published 30 Apr 2025
https://doi.org/10.35718/iiair.v1i1.1308

Grammar Correction: A Comparison of T5, LLAMA 2, and ChatGPT

Jihan Apriliani Nurhasanah[1], Healty Susantiningdyah[2], Iwan Saputra[3], Ilham Rahmaddani Adhie Prayoga[4], Muchammad Chandra Cahyo Utomo[5]

[1,3,4,5]Department of Mathematics and Information Technology, Institut Teknologi Kalimantan, Balikpapan, Indonesia
[2]Integrated Language Services Unit, Institut Teknologi Kalimantan, Balikpapan, Indonesia

Corresponding author: Muchammad Chandra Cahyo Utomo (ccahyo@lecturer.itk.ac.id)

To cite this article: J. A. Nurhasanah, H. Susantiningdyah, I. Saputra, I. Rahmaddani, A. Prayoga, M. C. C. Utomo, "Grammar Correction: A Comparison of T5, LLAMA 2, and ChatGPT," *Innovative Informatics and Artificial Intelligence Research*, vol. 1, issue 1, 2025. [Online]. Available: https://doi.org/10.35718/iiair.v1i1.1308

Abstract
English proficiency is a crucial tool for accessing new knowledge and skills and supporting self-directed learning across platforms and curricula. However, English language mastery in Indonesia has declined in recent years, as evidenced by decreasing rankings and scores compared to the Asian average and ASEAN countries. Grammatical errors significantly impact communication effectiveness, particularly in professional and academic environments that demand clarity and precision. To address this issue, AI-based Grammatical Error Correction (GEC) models offer a promising solution to enhance English learning outcomes. This study evaluates the performance of four GEC models: T5 Mini, T5 Tiny, LLAMA 2, and ChatGPT 3.5-turbo, focusing on their ability to detect and correct grammatical errors accurately and provide relevant feedback. The results show that LLAMA 2 achieves the best performance with the highest GLUE score (0.565), demonstrating its superiority in formal grammar correction tasks. T5 Mini follows with a score of 0.524, offering a balance between accuracy and efficiency. T5 Tiny, scoring 0.518, is suitable for resource-constrained environments despite its lower accuracy. ChatGPT 3.5-turbo, while having the lowest GLUE score (0.491), excels in providing cohesive and relevant feedback in conversational contexts. This research provides insights into the strengths and weaknesses of each model, aiding in the selection of the best solution to support automated English grammar learning.

Keywords: Grammatical Error Correction, LLAMA 2, T5, ChatGPT, Artificial Intelligence, Deep Learning, Large Language Model

1. Introduction
English is an essential tool for accessing new knowledge and skills, supporting deep self-directed learning across platforms and curricula. Whether as a first or second language, literacy is a vital asset for a nation to thrive in the digital world. However, over the past five years, English proficiency in Indonesia has shown a decline. In 2019, Indonesia ranked 61st, with a score dropping from 51.58 in 2018 to 50.06. This placed Indonesia below the average English proficiency score in Asia (53.00) and significantly behind other ASEAN countries such as Singapore, the Philippines, and Malaysia [1, 2].

Speaking skills are often regarded as the primary indicator of English proficiency because they enable direct communication and are easily observable. Therefore, English educators must design assessment strategies that provide constructive feedback on students' speaking abilities. Although many learning applications, both paid and free, are recommended to improve speaking skills, their usage often receives a lukewarm response. This is due to a lack of motivation and the absence of reward systems connected to classroom assessments.

Grammatical errors can hinder effective communication and have serious consequences, especially in professional and academic settings where clarity and accuracy are crucial. Grammar errors can also affect the writer's credibility and create confusion for readers. In recent years, the development of deep learning models for grammar error detection and correction has become an increasingly important area of research [3].

Advances in artificial intelligence technology provide significant opportunities to enhance English-speaking skill evaluation applications. One approach involves utilizing state-of-the-art language models to understand conversational contexts and deliver more accurate and natural feedback [4]. These models should be capable of generating text outputs in various contexts to produce the expected level of accurate and natural feedback.

However, when developing applications for evaluating speaking skills, it is important to consider different models with their respective advanta17 ges and disadvantages. The T5 Mini and T5 Tiny models, being smaller than the standard T5 model, allow for faster and more efficient operation on resource-constrained devices, albeit with slight compromises in accuracy compared to larger models [5].

On the other hand, LLAMA 2, developed by Meta, offers a different approach by focusing on processing efficiency and larger model sizes, which can provide better results in understanding more complex conversations. ChatGPT, developed by OpenAI, is recognized for its ability to generate natural conversations and deliver contextually relevant feedback with high engagement. The strength of ChatGPT lies in its ability to handle dynamic conversations and provide more human-like responses [6].

This study will compare the performance of several models, including T5 Mini, T5 Tiny, LLAMA 2, and ChatGPT, in the context of evaluating English-speaking skills. The main focus of this research is to assess the effectiveness of each model in providing accurate and relevant feedback, as well as their ability to detect and correct grammatical errors effectively [7].

The comparison will involve evaluating each model's performance in understanding conversational contexts and suggesting grammar corrections. The primary focus is on the accuracy of error detection and correction. Through this study, differences in the ability of each model to handle grammar errors in real-time and how this impacts the user learning experience will be examined.

Additionally, by testing various AI models, this study aims to identify the most efficient and effective solution for helping users improve their grammar skills. The results of this comparison are expected to provide deeper insights into which model excels in supporting English learning, particularly in the context of automated grammar correction, through a systematic and measurable evaluation of each model's capabilities. This research seeks to determine models that not only achieve high correction accuracy but also provide relevant and contextual improvement suggestions. Insights from this study will help identify the most efficient and effective solution for enhancing users' grammar skills.

2. Related Works

The development of methods for Grammatical Error Correction (GEC) has undergone significant advancements, starting from rule-based approaches to the widespread adoption of Large Language Models (LLMs) for GEC.

2.1 Rule-Based Method

Grammar correction methods have undergone significant development over time. Initially, these methods relied on rule-based approaches as their primary technique. Utilizing linguistic knowledge, grammar checks were conducted manually, such as verifying the subject-predicate structure of a sentence or whether it adhered to applicable grammatical rules. This approach was effective for addressing clear-cut errors but lacked flexibility and was limited in handling errors in more complex structures. An example of this is how Microsoft Word provides grammar correction suggestions to users [8].

2.2 Statistics

Statistical methods leverage probabilistic models, such as n-gram models, to detect anomalies based on word frequency patterns in large corpora. These methods are considered more effective than rule-based approaches but still have limitations, such as the need for vast amounts of data and the inability to capture the context of a sentence in a more complex manner [8].

2.3 *Machine Learning*

The advent of machine learning-based approaches brought significant improvements to grammar correction. These approaches reduce the need for manually crafting rules. However, they still require well-designed feature engineering to achieve optimal performance [8].

2.4 LLM

Large Language Models (LLMs) represent the most advanced approach to grammar correction today. Models like T5, LLAMA 2, and ChatGPT use Transformer-based architectures, enabling a deeper understanding of sentence context. Trained on massive corpora, LLMs can recognize complex error patterns and provide more natural correction suggestions [8].

3. Experimental Settings

The development of methods for Grammatical Error Correction (GEC) has undergone significant advancements, starting from rule-based approaches to the widespread adoption of Large Language Models (LLMs) for GEC.

This study utilizes the JFLEG dataset, which comprises input sentences (grammatically incorrect sentences) and reference corrections (manual corrections by humans). The dataset is processed to separate uncorrected sentences with up to four versions of manual corrections as references. The data is then formatted to be compatible with the models and evaluation metrics.

Three main models are tested in this research: ChatGPT, LLAMA 2, and T5. ChatGPT uses the 3.5 Turbo version accessed via the OpenAI API to generate automatic corrections, while LLAMA 2 is operated using specific prompts designed for grammar correction tasks. The T5 model is tested in two versions, T5 Mini and T5 Tiny, to evaluate their differences in implementation.

Each sentence in the JFLEG dataset is input into each model, and the resulting automatic corrections are saved. To evaluate performance, the GLEU metric is employed. GLEU compares the model's correction output to human references based on three key aspects: grammaticality, fluency, and meaning preservation. The input for GLEU includes the model's predicted sentences, manual correction references, and the original grammatically incorrect sentence for context. GLEU scores are calculated for each sentence and then averaged for each model. A comparison of GLEU scores across models is conducted to identify the best-performing model.

In addition to GLEU evaluation, the study includes qualitative evaluation by involving linguistic experts. In this evaluation, experts assess the corrections from each model by reviewing a paragraph of English text containing grammatical errors and comparing it to the model's corrections. This assessment considers aspects such as accuracy and contextual appropriateness. The qualitative evaluation provides deeper insights into the strengths and weaknesses of each model.

The test text used for this evaluation is a paragraph intentionally embedded with grammatical errors, as follows:

"*People always thinking about how to make the better world. But many peoples don't know where to begin or what to do. Education are important for peoples because it helps them become smart. But some children no have access to good schools. Another problems is pollution. Air and water becoming dirty everyday. This cause health problem to peoples and animals. We also must care about forest. Trees is cut down fast, and animals lose homes. It make the nature less balance. If everyone do small thing, the world can be more better. Even small actions like recycling or plant tree is helpful.*

Peoples need to work together if want solve big problems. But sometime they fight instead of help each other. This make everything more hard and slow. Government should do more, but some leaders no care about nature or peoples. They only think about money and power, so many problem still no fix.

Technology can be good thing, but it also create trouble. More factories mean more pollution, and peoples use too much plastic. Plastic never go away, it stay in ocean and hurt fish. We throw garbage anywhere and think it disappear, but it just move to another place and make problem there.

Many peoples think only big change matter, but that no true. If everyone stop wasting water or electric, it already big help. Using bike instead of car can also make less pollution. But peoples say it too hard or no time, so they continue bad habits. Change need effort, but it worth it.

Animals also important for balance of nature. If too many species disappear, ecosystem break. Peoples hunting animals for fun or for money, but they no think about future. Every small thing we do, like stop buying products from endangered animals, can save them. Nature give us life, so we must protect it."

Based on the text, it was subsequently corrected by grammar experts into the following:

"People are always thinking about how to make the world better, but many people don't know where to begin or what to do. Education is important for people because it helps them become smart. However, some children do not have access to good schools. Another problem is pollution. The air and water are becoming dirty every day. This causes health problems for people and animals. We must also care about the forests. Trees are being cut down quickly, and animals are losing their homes. This makes nature less balanced. If everyone does small things, the world can be better. Even small actions like recycling or planting trees are helpful.

People need to work together if they want to solve big problems. Nonetheless, sometimes they fight instead of helping each other. This makes everything more difficult and slow. The government should do more, but some leaders do not care about nature or people. They only think about money and power, so many problems remain unresolved.

Technology can be a good thing, but it also creates trouble. More factories mean more pollution, and people use too much plastic. Plastic never goes away; it stays in the ocean and hurts fish. We throw garbage anywhere and think it disappears, but it just moves to another place and creates problems there.

Many people think only big changes matter, but that's not true. If everyone stops wasting water or electricity, it can already be a big help. Riding a bike instead of a car can also create less pollution. However, people say it's too hard or they have no time, so they continue with their bad habits. Change needs effort, but it's worth it.

Animals are also important for the balance of nature. If too many species disappear, the ecosystem deteriorates. People hunt animals for fun or for money, but they do not think about the future. Every small thing we do, like stopping buying products made from endangered animals, can save them. Nature gives us life, so we must protect it."
This evaluation helps provide a better understanding of the models' capabilities in correcting grammatical errors.

4. Methods

4.1. Text-to-Text Transfer Transformer (T5)

The T5 model, or Text-to-Text Transfer Transformer, is a transformer-based architecture designed for natural language processing tasks using a text-to-text approach [9]. The T5 model has been widely used for grammatical error correction [10, 11, 12, 13, 14]. T5 employs stacks of self-attention layers in both its encoder and decoder to handle variable-length input. The encoder consists of self-attention layers and small feed-forward networks, enhanced by layer normalization and residual skip connections to improve processing efficiency. The decoder operates similarly but includes an additional attention mechanism that focuses on the encoder's output and generates vocabulary probabilities through a dense layer with a softmax function.

This study utilizes two variants of the T5 model: Base and Small. The Base model is the standard version with approximately 220 million parameters, while the Small version is a lighter model with 60 million parameters, featuring eight attention heads and six layers in both the encoder and decoder. These models were selected to balance prediction quality and computational efficiency, aligning with the requirements of the study.

4.2. LLAMA 2

LLAMA 2 is a series of large language models (LLMs) with parameter sizes ranging from 7 billion to 70 billion, designed to support dialogue applications. It includes a fine-tuned variant called LLAMA 2-Chat, optimized for conversational contexts [15]. Recent studies have leveraged LLAMA-based LLMs to address grammatical error correction (GEC) tasks in low-resource languages, demonstrating their effectiveness as both correction tools and generators of synthetic data. This is particularly evident in languages such as Estonian, Ukrainian, and German. In these applications, LLAMA 2 has shown promising results by producing high-quality data and outperforming open-source chat models in benchmarks of utility and safety [16].

4.3. Chat Generative Pre-Trained Transformer (ChatGPT)

ChatGPT is an artificial intelligence tool developed by OpenAI to generate text based on user input [17]. This model is designed to understand natural language and provide relevant, intelligent responses. ChatGPT is trained using a large amount of data to improve its ability to comprehend context and produce accurate answers. Research on the application of ChatGPT in Grammatical Error Correction (GEC) tasks shows that while its performance is lower compared to automatic evaluation metrics, ChatGPT has a unique ability to perform more comprehensive corrections. This model does not only correct errors word by word but can also change surface expressions and sentence structures while maintaining grammatical accuracy. Human evaluation confirms that ChatGPT produces fewer under-corrections and mis-corrections, although it generates more over-corrections [18]. This emphasizes ChatGPT's potential as a useful tool for grammatical error correction, as well as for enhancing the quality of expressions and sentence structures in a holistic manner.

4.3 Generalized Language Evaluation Understanding (GLEU)

Grammatical Error Correction Evaluation Metric (GLEU) is a metric developed to better assess grammatical error correction (GEC) tasks compared to existing GEC metrics. GLEU is inspired by the BLEU metric used in machine translation. The goal of this metric is to address the issue of the lack of a clear "ground truth" or reference in GEC evaluation, which has traditionally relied on the correlation of corrected errors with references or intuition. GLEU more accurately reflects human judgment regarding the quality of grammatical error correction systems because it aligns better with the rankings produced from human evaluations in GEC tasks, such as those conducted in the CoNLL-2014 Shared Task [19].

$$Count_B(n-gram) = \sum_{n-gram' \in B} d(n-gram, n-gram') \quad (1)$$

The n-gram calculation is obtained from Equation 1. The n-gram calculation for model's prediction is shown in Equation 1. The function $d(n-gram, n-gram')$ represents whether n-gram is appropriate or not. To ensure only the relevant n-gram being calculated, the Equation 2 is used.

$$d(n-gram, n-gram') = \begin{cases} 1 & \text{if } n-gram = n-gram' \\ 0 & otherwise \end{cases} \quad (2)$$

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-c/r} & c \leq r \end{cases} \quad (3)$$

A brevity penalty (BP) is used to penalize predictions that are too short (i.e., do not cover all the necessary context) as shown in Equation 3. If the prediction is longer than the reference, no penalty is applied. If the prediction is shorter, an exponential penalty is applied to reduce the score.

$$GLEU(C,R,S) = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p'_n\right) \quad (4)$$

The main formula for combining the n-gram probabilities of the model's prediction (p'_n) for all n-grams (from unigram to N-gram) is shown as in Equation 4. The weight for each n-gram is denoted by w_n. The probability of the n-th n-gram is calculated as the ratio between the number of n-grams that match in the prediction and the reference to the total n-grams in the prediction, denoted as p'_n.

This formula is very similar to BLEU but has a different focus. BLEU is used for machine translation evaluation, comparing the model's output only with human references. On the other hand, GLEU is used for grammatical error correction evaluation, considering the original input, the model's output, and human references. It penalizes predictions that are too similar to the original input without substantial improvements.

4.4 JHU Fluency-Extended GUG corpus (JFLEG)

JFLEG (The Johns Hopkins Fluency-Enhanced Grammar Error Correction) is a dataset consisting of texts with various levels of language proficiency, where each entry includes the original sentence and its corrected version. The corrections in the dataset employ a holistic fluency approach, which not only corrects grammatical errors but also enhances the naturalness of the text. JFLEG provides various types of corrections, allowing both qualitative and quantitative analysis of the performance of four leading GEC systems. Evaluation results show significant differences in system performance when using a fluency corpus compared to a minimal-edit corpus, and highlight errors that are often overlooked, such as spelling and long-distance context errors. Therefore, JFLEG serves as a new standard for more accurate and effective GEC evaluation [4].

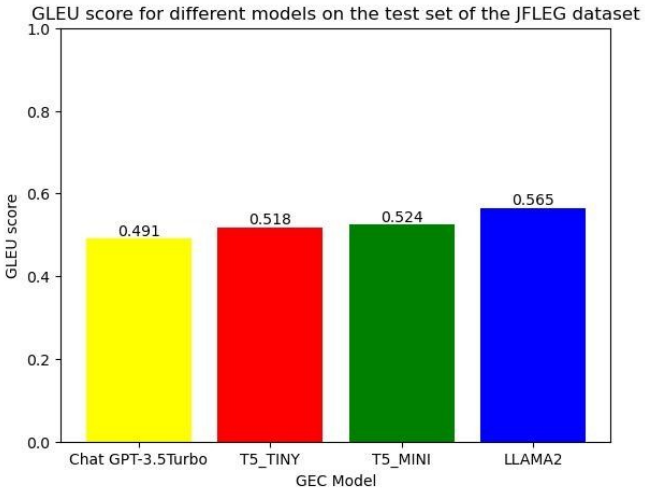

Figure 1: Model's evaluation based on GLEU Score

5. Results and Discussions

The models used in this evaluation include Chat-GPT 3.5-turbo via API, T5_Tiny, T5_Mini, and LLAMA 2. Chat-GPT 3.5-turbo was accessed through the API provided by OpenAI, allowing seamless integration to generate high-quality and contextually relevant responses. T5_Tiny and T5_Mini are variants of the T5 model, with T5_Tiny designed as a smaller and more efficient model, while T5_Mini offers medium-sized performance that surpasses the smallest version. LLAMA 2, developed by Meta, is a large language model

with the capability to generate highly relevant and high-quality text.

The evaluation was conducted by selecting the maximum value from each prediction generated by these models, and the scores were averaged to compute the evaluation metric using GLEU. The results of the model evaluation are presented in Figure 1.

Based on Figure 1, the comparison of GLEU scores obtained from the four Grammatical Error Correction (GEC) models tested using the JFLEG test dataset is presented. The evaluated models include ChatGPT 3.5-turbo, T5 Tiny, T5 Mini, and LLAMA 2.

From the evaluation results, LLAMA 2 demonstrates the best performance with the highest GLEU score of 0.565, indicating its superior ability to generate grammatical corrections that closely align with the manual references.

The second-best performing model is T5 Mini, achieving a GLEU score of 0.524, slightly behind LLAMA 2. This suggests that T5 Mini is quite capable of maintaining accuracy, fluency, and semantic consistency in the corrected sentences.

Following T5 Mini is T5 Tiny, ranking third with a GLEU score of 0.518. Although its score is slightly lower than T5 Mini, this result indicates that T5 Tiny can still provide corrections that closely approximate the references, despite being a smaller model in terms of parameter size compared to T5 Mini.

In the last position is ChatGPT 3.5-turbo, which scored 0.491, the lowest among the four models. This score suggests that ChatGPT encounters greater challenges in producing grammatical corrections that match human references, particularly when handling more complex sentence structures. However, this result may also be influenced by ChatGPT's optimization for conversational contexts rather than pure text correction tasks like those in JFLEG.

To further analyze the effectiveness of each model, an example paragraph was used as a demonstration. The grammatical errors in the text will be corrected directly using the developed models. The original text used is as follows:

"*People always thinking about how to make the better world. But many peoples don't know where to begin or what to do. Education are important for peoples because it helps them become smart. But some children no have access to good schools. Another problems is pollution. Air and water becoming dirty everyday. This cause health problem to peoples and animals. We also must care about forest. Trees is cut down fast, and animals lose homes. It make the nature less balance. If everyone do small thing, the world can be more better. Even small actions like recycling or plant tree is helpful.*

Peoples need to work together if want solve big problems. But sometime they fight instead of help each other. This make everything more hard and slow. Government should do more, but some leaders no care about nature or peoples. They only think about money and power, so many problem still no fix.

Technology can be good thing, but it also create trouble. More factories mean more pollution, and peoples use too much plastic. Plastic never go away, it stay in ocean and hurt fish. We throw garbage anywhere and think it disappear, but it just move to another place and make problem there.

Many peoples think only big change matter, but that no true. If everyone stop wasting water or electric, it already big help. Using bike instead of car can also make less pollution. But peoples say it too hard or no time, so they continue bad habits. Change need effort, but it worth it.

Animals also important for balance of nature. If too many species disappear, ecosystem break. Peoples hunting animals for fun or for money, but they no think about future. Every small thing we do, like stop buying products from endangered animals, can save them. Nature give us life, so we must protect it. "

The results of the tested models are presented as shown in Table 1.

Table 1: Comparison Table of Correction Results

Model	Teks
Dataset	*People **are** always thinking about how to make the **world better**, but many **people** don't know where to begin or what to do. Education **is** important for **people** because it helps them become smart. **However**, some children **do not** have access to good schools. Another **problem** is pollution. **The air** and water **are** becoming dirty **every day**. This **causes** health **problems for people** and animals. We **must also** care about **the forests**. **Trees are being** cut down **quickly**, and animals **are losing their homes**. **This makes** nature less **balanced**. If everyone **does** small **things**, the world can be better. Even small actions like recycling or **planting trees are** helpful.* ***People** need to work together if **they want to** solve big problems. **Nonetheless, sometimes** they fight instead of **helping** each other. This **makes** everything more **difficult** and slow. **The government** should do more, but some leaders **do not** care about nature or **people**. They only think about money and power, so many **problems remain unresolved**.* *Technology can be **a** good thing, but it also **creates** trouble. More factories mean more pollution, and **people** use too much plastic. Plastic never **goes away**; it **stays** in **the** ocean and **hurts** fish. We throw garbage anywhere and think it **disappears**, but it just **moves** to another place and **creates problems** there.* *Many **people** think only big **changes** matter, but **that's not** true. If everyone **stops** wasting water or **electricity**, it **can already be a** big help. **Riding a** bike instead of **a** car can also **create** less pollution. **However, people** say **it's** too hard or **they have** no time, so they continue **with their** bad habits. Change **needs** effort, but **it's** worth it.* *Animals **are** also important for **the** balance of nature. If too many species disappear, **the***

	ecosystem deteriorates. ***People hunt*** *animals for fun or for money, but they* ***do not*** *think about* ***the*** *future. Every small thing we do, like* ***stopping*** *buying products* ***made*** *from endangered animals, can save them. Nature* ***gives*** *us life, so we must protect it.*
ChatGPT	*People* ***are*** *always thinking about how to make the* ***world better***. *But many people don't know where to begin or what to do. Education* ***is*** *important for* ***people*** *because it helps them become smart. But some children* ***do not*** *have access to good schools. Another* ***problem*** *is pollution.* ***The air*** *and water* ***are*** *becoming dirty* ***every day***. *This* ***causes*** *health* ***problems for people*** *and animals. We* ***must*** *also care about* ***the forests***. ***Trees are being*** *cut down* ***quickly***, *and animals* ***are losing their homes***. ***This makes*** *nature less* ***balanced***. *If everyone* ***does*** *small* ***things***, *the world can be better. Even small actions like recycling or* ***planting trees are*** *helpful.* ***People*** *need to work together if* ***they want to*** *solve big problems. But* ***sometimes*** *they fight instead of* ***helping*** *each other. This* ***makes*** *everything more* ***difficult*** *and slow.* ***The government*** *should do more, but some leaders* ***do not*** *care about nature or* ***people***. *They only think about money and power, so many* ***problems remain unresolved***. *Technology can be* ***a*** *good thing, but it also* ***creates*** *trouble. More factories mean more pollution, and* ***people*** *use too much plastic. Plastic never* ***goes away***; *it* ***stays*** *in* ***the*** *ocean and* ***hurts*** *fish. We throw garbage anywhere and think it* ***disappears***, *but it just* ***moves*** *to another place and* ***creates problems*** *there.* *Many* ***people*** *think only big* ***changes*** *matter, but* ***that's not*** *true. If everyone* ***stops*** *wasting water or* ***electricity***, *it* ***can already be a*** *big help. Using* ***a*** *bike instead of* ***a*** *car can also* ***create*** *less pollution. But* ***people*** *say* ***it's*** *too hard or* ***they have*** *no time, so they continue* ***with their*** *bad habits. Change* ***needs effort***, *but* ***it's*** *worth it.* *Animals* ***are*** *also important for* ***the*** *balance of nature. If too many species disappear,* ***the*** *ecosystem* ***breaks***. ***People hunt*** *animals for fun or for money, but they* ***do not*** *think about* ***the*** *future. Every small thing we do, like* ***stopping*** *buying products from endangered animals, can save them. Nature* ***gives*** *us life, so we must protect it.*
T5 Tiny	*People always* ***think*** *about how to make a better world. But many* ***people*** *don't know where to begin or what to do. Education* ***is*** *important for* ***people*** *because it helps them become smart. But some children* ***have no*** *access to good schools. Another* ***problem*** *is pollution. Air and water* ***become*** *dirty* ***every day***. ***This causes a*** *health problem* ***for people*** *and animals. We also must care about* ***the*** *forest. Trees* ***are*** *cut down fast, and animals lose homes. It* ***makes*** *nature less balance. If everyone* ***does a*** *small thing, the world can be better. Even small actions like recycling or* ***planting trees are*** *helpful.* ***People*** *need to work together if* ***they want to*** *solve big problems. But* ***sometimes*** *they fight instead of* ***helping*** *each other. This* ***makes*** *everything more hard and slow.* ***The government*** *should do more, but some leaders* ***have*** *no care about nature or* ***people***. *They only think about money and power, so many* ***problems*** *still* ***have*** *no fix.* *Technology can be* ***a*** *good thing, but it also* ***creates*** *trouble. More factories mean more pollution, and* ***people*** *use too much plastic.* ***Plastics*** *never go away,* ***they*** *stay in* ***the*** *ocean and hurt fish. We throw garbage anywhere and think it* ***disappears***, *but it just* ***moves*** *to another place and* ***makes problems there*** *Many* ***people*** *think only* ***a*** *big change matter, but that* ***is*** *no true. If everyone* ***stops*** *wasting water or* ***electricity***, *it* ***is*** *already* ***a*** *big help. Using* ***a*** *bike instead of* ***a*** *car can also make less pollution. But* ***people*** *say it* ***is*** *too hard or no time, so they continue bad habits. Change* ***needs an*** *effort, but it* ***is*** *worth it.* *Animals* ***are*** *also important for* ***the*** *balance of nature. If too many species disappear,* ***the ecosystem break***s. ***People*** *hunting animals for fun or for money, but they* ***don't*** *think about* ***the*** *future. Every small thing we do, like* ***stopping*** *buying products from endangered animals, can save them. Nature* ***gives*** *us* ***a*** *life, so we must protect it.*
T5 Mini	*People always* ***think*** *about how to make the* ***world better***. *But many* ***people*** *don't know where to begin or what to do. Education* ***is*** *important for* ***people***, *because it helps them become smart. But some children* ***have no*** *access to good schools. Another* ***problem*** *is pollution. Air and water* ***are*** *becoming dirty* ***every day***. ***This causes a*** *health problem* ***for people*** *and animals. We also must care about* ***the*** *forest. Trees are cut down fast, and animals lose* ***their*** *homes. It* ***makes*** *nature less balance. If everyone* ***does a*** *small thing, the world can be better. Even small actions like recycling or* ***planting trees are*** *helpful.* ***People*** *need to work together if* ***they want to*** *solve big problems. But* ***sometimes*** *they fight instead of* ***helping*** *each other. This* ***makes***

	*everything more hard and slow. **The government** should do more, but some leaders do no care about nature or **people**. They only think about money and power, so many **problems** still **have** no fix.* *Technology can be **a** good thing, but it also **causes** trouble. More factories mean more pollution, and **people** use too much plastic. Plastic never **goes** away, it **stays** in **the ocean** and hurts fish. We throw garbage anywhere and think it **disappears**, but it just **moves** to another place and **causes problems** there.* *Many **people** think only **about** big **changes in the** matter, but that **is not** true. If everyone **stops** wasting water or **electricity**, it **is** already **a** big help. Using **a** bike instead of **a** car can also **cause** less pollution. But **people** say it **is** too hard or **not a** time, so they continue bad habits. Change **needs an** effort, but it **is** worth it.* *Animals **are** also important for **the** balance of nature. If too many species **disappears**, **the ecosystem breaks down**. **People** hunting animals for fun or for money, but they **don't** think about **the** future. Every small thing we do, like stop buying products from endangered animals, can save them. Nature **gives** us **a** life, so we must protect **it**.*
LLAMA 2	*People **are** always thinking about how to make a better world. But many **people** don't know where to begin or what to do. Education **is** important for **people** because it helps them become smart. **However**, some children **don't** have access to good schools. Another **problem** is pollution. Air and water **are** becoming dirty **every day**. This **causes** health **problems for people** and animals. We also must care about **the** forest. Trees **are being** cut down **too** fast, and animals **are losing their** homes. It **makes** the **natural balance** less **stable**. If everyone **does** small **things**, the world can be more better. Even small actions like recycling or **planting trees are** helpful.* ***People** need to work together if **they want to** solve big problems. **However**, **sometimes** they fight instead of **helping** each other. This **makes** everything more **difficult** and slow. **The government** should do more, but some leaders **don't** care about nature or **people**. They only think about money and power, so many **problems still need to be fixed**.* *Technology can be **a** good thing, but it also **creates troubles**. More factories mean more pollution, and **people** use too much plastic. Plastic never **goes** away, it **stays** in **the** ocean and **hurts** fish. We throw garbage anywhere and think it **disappears**, but it just **moves** to another place and **makes problems** there.* *Many **people** think **that** only big **changes** matter, but that **is not** true. If everyone **stopped** wasting water or **electricity**, it **would already be a** big help. Using **a** bike instead of **a** car can also **reduce pollution**. **However**, **people often** say it **is** too hard or **they don't have** time, so they continue **with their** bad habits. Change **requires** effort, but it **is** worth it.* *Animals **are** also important for **maintaining the** balance of nature. If too many species disappear, **the ecosystem can become unstable**. **People are** hunting animals for fun or for money, but they **don't** think about the **future consequences**. Every small thing we do, **such as** stop buying products **made** from endangered animals, can **help** save them. Nature **gives** us life, so we must protect it.*

Based on Table 1, it can be observed that the comparison of correction results from four models—ChatGPT, T5 Mini, T5 Tiny, and LLAMA 2—was conducted against grammatically correct text to evaluate the accuracy of each model.

The ChatGPT model demonstrates the best performance in grammar correction, particularly in creating natural, structured, and cohesive text. The use of words like "*However*" and "*Nonetheless*" adds a smooth transition between sentences, reflecting an understanding of how to logically connect ideas. Sentences such as "If everyone stops wasting water or electricity, it can already be a big help" showcase proper grammar usage, with clear subject-predicate structure and no tense errors. Additionally, the sentence structure includes not only simple forms but also complex ones, providing stylistic variety that makes the text more engaging to read. Throughout the text, no noticeable grammar errors were found, and the tone feels professional.

LLAMA 2 comes close to the quality of ChatGPT but exhibits some notable weaknesses in grammar and word choice. Phrases like "*the world can be more better*" highlight grammatical redundancy, as the use of both "more" and "*better*" is unnecessary. On the other hand, LLAMA 2 still demonstrates strength in managing more complex sentence structures compared to T5 Tiny and T5 Mini, with examples like "*Technology can be a good thing, but it also creates troubles.*" However, LLAMA 2 has some sentences that lack precision, such as "*so many problems still need to be fixed*," which feels less refined compared to ChatGPT's version. In formal contexts, this model requires slight improvements to produce flawless text.

T5 Tiny tends to generate simpler text, but this comes at the expense of grammatical accuracy. Examples of errors include "*It makes nature less balance*," where the adjective "*balance*" should be corrected to "balanced" to follow proper grammatical rules. This model also frequently produces short sentences with limited variation, such as "*Air and water become dirty every day*," which sounds overly simplistic and less polished than ChatGPT's version. Moreover, T5 Tiny often misses the nuance in more complex text, making its

output less suitable for formal use. While the model conveys the core message, its text tends to be less engaging due to the absence of natural transitions.

T5 Mini performs slightly better than T5 Tiny in terms of structure and grammar but still exhibits similar weaknesses. Grammar errors, such as "*It makes nature less balance,*" persist in this model, highlighting limitations in understanding the context of adjective-to-passive verb transformations. Nonetheless, T5 Mini provides smoother structures compared to T5 Tiny, with sentences like "*Air and water are becoming dirty every day*," which are closer to correct grammar. However, its sentences still feel too simple and lack dynamism. In transitioning between ideas, T5 Mini falls short of ChatGPT or LLAMA 2, resulting in text that feels flatter and less cohesive.

6. Conclusions

From the performance evaluation of the four GEC models on the JFLEG dataset, it can be concluded that each model has distinct strengths and weaknesses suited to different contexts of use. LLAMA 2 demonstrated the best performance with the highest GLEU score of 0.565, indicating its ability to produce grammatical corrections that closely align with human references, particularly for tasks based on formal datasets. However, despite its overall superiority, LLAMA 2 exhibits some weaknesses in certain word choices and sentence structures, making it less precise for formal applications.

Further studies could focus on developing learning applications using LLMs and exploring other models in terms of efficiency. Our study can serve as a reference for selecting suitable LLMs when developing English language applications, such as automatic IELTS assessment tools. Fine-tuning or identifying robust architectures for deep learning models is also worth investigating.

Declaration of Conflicting Interests

The authors declare that there are no competing interests that could have influenced the work of our study.

References

[1] EF Education First, "EF English Proficiency Index 2018," 2018, Retrieved from https://ef.com/assetscdn/WIBIwq6RdJvcD9bc8RMd/cefcom-epi-site/reports/2018/ef-epi-2018-english.pdf

[2] EF Education First, "EF English Proficiency Index 2019," 2019, Retrieved from https://www.ef.com/assetscdn/WIBIwq6RdJvcD9bc8RMd/cefcom-epi-site/reports/2019/ef-epi-2019-english.pdf

[3] H. A. Z. S. Shahgir and K. S. Sayeed, "Bangla Grammatical Error Detection Using T5 Transformer Model," 2023, arXiv. doi: 10.48550/ARXIV.2303.10612.

[4] F. Ahsan, "Grammatical Error Correction with Transformer Models - Scribendi AI," Scribendi AI, Feb. 04, 2021. https://www.scribendi.ai/grammatical-error-correction-with-transformer-models/ (accessed Jan. 3, 2025).

[5] B. Kim, K. Lee, J. Kim, and S. Lee, "Small Language Models are Equation Reasoners," 2024, arXiv. doi: 10.48550/ARXIV.2409.12393.

[6] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," 2023, arXiv. doi: 10.48550/ARXIV.2302.13971.

[7] C. Napoles, K. Sakaguchi, and J. Tetreault, 'JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction', in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2*, Short Papers, 2017, pp. 229–234.

[8] M. Naghshnejad, T. Joshi, and V. N. Nair, "Recent Trends in the Use of Deep Learning Models for Grammar Error Handling," 2020, arXiv. doi: 10.48550/ARXIV.2009.02358.

[9] Colin Raffel, undefined., et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *in J. Mach. Learn. Res.*, vol. 21, pp. 140:1-140:67, 2019.

[10] J., Matias. "Grammatical Error Correction with byte-level language models (Master's thesis), " Universitet I Oslo, 2023.

[11] A. Vadehra and P. Poupart, "Detecting Errors to Improve Grammar Error Correction Models - Scribendi AI," Scribendi AI, Jun. 29, 2023. https://www.scribendi.ai/detecting-errors-to-improve-grammar-error-correction-models/

[12] M. Harahus et al., "Evaluation of Datasets Focused on Grammatical Error Correction Using the T5 Model in Slovak," in *2024 34th International Conference Radioelektronika (RADIOELEKTRONIKA). IEEE*, pp. 1–6, Apr. 17, 2024. doi: 10.1109/radioelektronika61599.2024.10524071.

[13] A. Katinskaia and R. Yangarber, "Grammatical Error Correction for Sentence-level Assessment in Language Learning," in *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023). Association for Computational Linguistics*, pp. 488–502, 2023. doi: 10.18653/v1/2023.bea-1.41.

[14] M. Qorib, H. Ng, "Grammatical Error Correction: Are We There Yet?," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 2794–2800.

[15] H. Touvron et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," 2023, arXiv. doi: 10.48550/ARXIV.2307.09288.

[16] A. Luhtaru, T. Purason, M. Vainikko, M. Del, and M. Fishel, "To Err Is Human, but Llamas Can Learn It Too," in *Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics*, pp. 12466–12481, 2024. doi: 10.18653/v1/2024.findings-emnlp.727.

[17] T. M. Sahib et al., "A comparison between ChatGPT-3.5 and ChatGPT-4.0 as a tool for paraphrasing English Paragraphs". In *Int. Applied Social Sciences*, pp. 471-

480, 2023

[18] H. Wu, W. Wang, Y. Wan, W. Jiao, and M. Lyu, “ChatGPT or Grammarly? Evaluating ChatGPT on Grammatical Error Correction Benchmark,” 2023, arXiv. doi: 10.48550/ARXIV.2303.13648.

[19] C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault, “Ground Truth for Grammaticality 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)*. *Association for Computational Linguistics*, 2015. doi: 10.3115/v1/p15-2097.