Targeted Syntactic Evaluation for Grammatical Error Correction

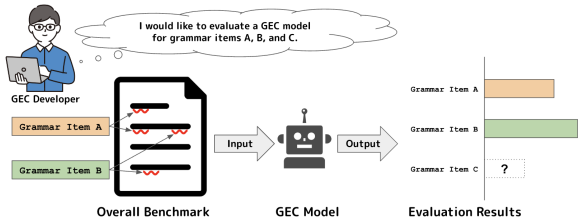
Aomi Koyama[1] **Masato Mita**[1,2] **Su-Youn Yoon**[3,2]
Yasufumi Takama[1] **Mamoru Komachi**[2,1]
[1]Tokyo Metropolitan University [2]Hitotsubashi University [3]EduLab, Inc.
koyama-aomi@ed.tmu.ac.jp, mita@edu.sds.hit-u.ac.jp, su-youn.yoon@edulab-inc.com
ytakama@tmu.ac.jp, mamoru.komachi@r.hit-u.ac.jp

Abstract

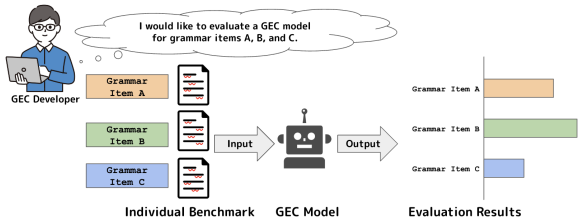
Language learners encounter a wide range of grammar items across the beginner, intermediate, and advanced levels. To develop grammatical error correction (GEC) models effectively, it is crucial to identify which grammar items are easier or more challenging for models to correct. However, conventional benchmarks based on learner-produced texts are insufficient for conducting detailed evaluations of GEC model performance across a wide range of grammar items due to biases in their distribution. To address this issue, we propose a new evaluation paradigm that assesses GEC models using minimal pairs of ungrammatical and grammatical sentences for each grammar item. As the first benchmark within this paradigm, we introduce the **C**EFR-based **T**argeted **S**yntactic **E**valuation Dataset for **G**rammatical Error Correction (CTSEG), which complements existing English benchmarks by enabling fine-grained analyses previously unattainable with conventional datasets. Using CTSEG, we evaluate three mainstream types of English GEC models: sequence-to-sequence models, sequence tagging models, and prompt-based models. The results indicate that while current models perform well on beginner-level grammar items, their performance deteriorates substantially for intermediate and advanced items.

1 Introduction

Grammatical error correction (GEC) is a task that involves correcting errors in text to produce grammatically correct expressions. It primarily aims to support language learners by addressing various types of errors they make (Bryant et al., 2023). Currently, sequence-to-sequence models (Kiyono et al., 2020; Rothe et al., 2021) and sequence tagging models (Omelianchuk et al., 2020; Tarnavskyi et al., 2022) are widely used in GEC, both of which have achieved high performance. In recent years, models leveraging instruction-based prompts (Brown et al., 2020; OpenAI et al., 2024) have also gained significant attention in this field, demonstrating promising results (Wu et al., 2023; Fang et al., 2023; Coyne et al., 2023; Loem et al., 2023; Davis et al., 2024).



(a) Conventional evaluation datasets based on learner-produced text: Since these datasets are not explicitly designed for specific grammar items, they may not cover all the items that developers intend to evaluate.



(b) A proposed evaluation dataset: Developers create dedicated evaluation datasets for specific grammar items, enabling targeted assessment of GEC models.

Figure 1: Comparison of evaluation methods: Conventional vs. proposed evaluation dataset for GEC models. Red wavy lines indicate errors for each grammar item.

In GEC, learner-generated texts have traditionally been used for model evaluation. The CoNLL-2014 shared task test set (Ng et al., 2014), one of the most widely recognized benchmarks in GEC, is based on essays written by students at the National University of Singapore. Similarly, the JFLEG test set (Napoles et al., 2017), which focuses on measuring correction fluency (Sakaguchi et al., 2016), includes essays written by learners at various proficiency levels as part of English proficiency tests. These test sets enable GEC models to be evaluated in contexts that closely reflect real-world usage.

However, evaluating GEC models using learner-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21108–21125

Dataset	Sents.	Refs.	Error Tags	CEFR Level
FCE (Yannakoudakis et al., 2011)	2,695	1	71	B1–B2
KJ (Nagata et al., 2011)	3,199	1	22	A1–A2?
CoNLL-2014 (Ng et al., 2014)	1,312	2	28	C1
AESW (Daudaravicius et al., 2016)	143,804	1	N/A	C1–C2 (+Native)
JFLEG (Napoles et al., 2017)	747	4	N/A	A1–C2?
BEA-2019 (Bryant et al., 2019)	4,384	5	25	A1–C2 (+Native)
GMEG (Napoles et al., 2019)	2,960	4	N/A	B1–B2 (+Native)
CWEB (Flachs et al., 2020)	6,845	2	25	(Native)
CTSEG (Ours)	1,578	1–3	263	A1–B2

Table 1: A summary of English GEC evaluation datasets and other relevant datasets. A question mark (?) in the CEFR level column indicates unknown or approximated information.

produced texts presents several challenges. For instance, the range of grammar items that can be assessed depends on the content of learners' essays. As a result, certain grammar items that developers wish to examine (e.g., grammar item C in Figure 1a) may be absent, leaving the model's performance on those items unknown. From the perspective of learner support, this is undesirable, as it may lead to certain grammar items being introduced to learners with suboptimal correction accuracy. Moreover, when a single sentence contains multiple errors, the corrections may interact with one another (Mita and Yanaka, 2021), making it difficult to isolate and accurately assess the model's performance on specific grammar items.

To address these issues, we propose integrating Targeted Syntactic Evaluation (TSE) into the GEC evaluation framework (Figure 1b). TSE assesses grammatical and syntactic knowledge by comparing the probabilities assigned to minimal pairs of acceptable and unacceptable sentences, enabling fine-grained performance evaluation across specific linguistic phenomena (Linzen et al., 2016; Marvin and Linzen, 2018; Warstadt et al., 2020; Someya et al., 2024). Although TSE has traditionally been used for evaluating language models, its application to downstream tasks such as GEC is not straightforward. However, unlike other generative tasks, such as machine translation and summarization, GEC inherently aligns with the minimal pair setting, where only specific parts of a sentence are modified rather than the entire sentence. This structural similarity enables the adaptation of the TSE concept for GEC evaluation.

Building on this framework, we introduce the **C**EFR-based **T**argeted **S**yntactic **E**valuation Dataset for **G**rammatical Error Correction (CTSEG), the first dataset developed under this paradigm, designed to facilitate proficiency-aligned evaluations of GEC models.[1] Grounded in the CEFR (Council of Europe, 2001), CTSEG ensures that evaluations align with learner proficiency levels, providing a more structured and pedagogically relevant assessment. By constructing minimal pairs for different grammar items, it enables a targeted analysis of model performance on specific grammatical structures. To assess the effectiveness of our dataset, we conducted extensive experiments on three major types of GEC approaches: sequence-to-sequence models, sequence tagging models, and prompt-based models. The results reveal strengths and weaknesses in how different approaches handle specific grammatical phenomena, highlighting limitations that conventional leaderboard-style evaluations fail to capture and underscoring the necessity of fine-grained assessment for advancing GEC research.

In addition, leveraging CTSEG, we conducted an experiment to explore the applicability of prompt-based models—an area of growing interest in recent years—for learner support. Specifically, we investigated whether such models could perform corrections using a target grammar item specified in the prompt, while varying the type of reference sentence used for evaluation. The results showed that the model was indeed capable of incorporating the specified grammar item into its corrections. This suggests a promising feature for learners who wish to focus on practicing specific grammar items.

2 Related Work

2.1 GEC Evaluation

Various evaluation datasets have been published for GEC research. Table 1 lists evaluation datasets for GEC and related resources. The CoNLL-2014

[1]The CTSEG dataset is publicly available at https://github.com/SDS-NLP/CTSEG.

shared task test set (Ng et al., 2014) is one of the most widely used benchmarks in GEC and is based on the NUCLE corpus (Dahlmeier et al., 2013). The JFLEG test set (Napoles et al., 2017) was developed to assess the fluency-oriented performance of GEC models and is based on the GUG corpus (Heilman et al., 2014). The BEA-2019 shared task test set (Bryant et al., 2019) is derived from the W&I and LOCNESS corpora (Yannakoudakis et al., 2018; Granger, 1998) and includes both learner-written texts categorized by CEFR levels and essays written by native speakers. All these datasets, as well as related resources, are based on essays written by learners or native speakers. Therefore, the types of errors present in these evaluation datasets are not systematically controlled. As a result, these datasets do not necessarily include the specific types of errors that developers need to test in GEC models. To address this issue, we propose the creation of a controlled evaluation dataset designed to assess GEC models on specific error types.

In GEC evaluation, error tags in evaluation datasets can be used to assess model performance across different error types. The CoNLL-2014 test set includes 28 types of error tags, primarily focusing on common learner errors such as tense and article usage. The BEA-2019 test set employs 25 types of error tags based on the ERRANT (Felice et al., 2016; Bryant et al., 2017)[2] annotation scheme. However, since it relies on rule-based methods, its error tags are restricted to errors that can be classified using predefined rules. As a result, the error tags in the CoNLL-2014 and BEA-2019 test sets do not allow for a detailed performance analysis of rare errors or those that are difficult to categorize using rule-based methods. In contrast, our tags offer the advantage of providing detailed information about the target grammatical expressions associated with grammatical errors.[3] For instance, verb tense errors, which are categorized as a single type in both the CoNLL-2014 and BEA-2019 test sets, are further classified into 15 subcategories in our dataset, including present, present perfect, present progressive, and present perfect progressive. This granularity allows us to pinpoint specific grammatical patterns where GEC systems underperform. Thus, CTSEG enables a more precise measurement of performance for error types that are challenging to evaluate using existing datasets.

[2]ERRANT is a rule-based tool that automatically identifies error types and assigns corresponding error tags.

[3]A list of grammar items included in CTSEG can be found in Appendix A.

2.2 GEC Approaches

Recent GEC research primarily employs three approaches: sequence-to-sequence methods, sequence tagging methods, and prompt-based methods. Sequence-to-sequence methods (Kiyono et al., 2020; Rothe et al., 2021) utilize encoder-decoder architectures, such as Transformer (Vaswani et al., 2017), to generate corrected sentences. The introduction of sequence-to-sequence methods has substantially improved GEC model performance compared to earlier approaches based on language models (Gamon et al., 2008) or classifiers (Dahlmeier and Ng, 2011). Sequence tagging methods (Omelianchuk et al., 2020; Tarnavskyi et al., 2022) predict edit operation tags for input sentences, which are then applied to generate corrected outputs. Compared to sequence-to-sequence methods, sequence tagging methods tend to achieve higher precision but often exhibit lower recall (Omelianchuk et al., 2020). Prompt-based methods (Wu et al., 2023; Davis et al., 2024) leverage large language models, such as GPT-4 (OpenAI et al., 2024), to generate corrected sentences. Compared to other methods, prompt-based methods typically produce more natural and fluent corrections (Davis et al., 2024). Additionally, modifying the prompts allows for some degree of control over the corrections (Loem et al., 2023). In this study, we select and analyze a representative model from each of these three approaches, evaluating their performance using our newly created dataset.

3 The CTSEG Dataset

This section describes the methodology used to construct CTSEG dataset designed for grammar-item-specific evaluation.

3.1 Preliminary Preparation

This subsection provides an overview of the annotators involved in creating minimal pairs and presents the grammar item list used as evaluation criteria for GEC models.

Annotators We recruited two experienced English teachers to create minimal pairs of ungrammatical and grammatical sentences. Both teachers have over 10 years of experience teaching English. They possess extensive expertise in coaching English composition and grammar. These experienced educators were selected to ensure that the

ID	CL	Grammar Item	Sentence	Procedure
12	A1	Demonstrative adjective (these/those + noun)	*That pants are too big. Those pants are too big.	Step 2 Step 2
39	A2	Comparative (superiority) (more + adjective/adverb)	*This dictionary is most useful than the one you bought. This dictionary is more useful than the one you bought.	Step 2 Step 2
173	B1	Relative pronoun (nominative) (which)	*I am looking for a game what will amuse my children. I am looking for a game which will amuse my children. I am looking for a game that will amuse my children.	Step 2 Step 2 Step 3
175	B2	Relative pronoun (objective) (who)	*The person which I met yesterday is a doctor. The person who I met yesterday is a doctor. The person __ I met yesterday is a doctor. The person that I met yesterday is a doctor.	Step 2 Step 2 Step 3 Step 3

Table 2: Examples of minimal pairs for each grammar item. The ID column represents a unique number assigned to each grammar item. The CL column specifies the CEFR level associated with each grammar item. An asterisk (*) denotes an ungrammatical sentence.

ungrammatical expressions in the sentences accurately reflect common student errors. Additionally, involving multiple annotators in the sentence creation process helps mitigate potential biases in error representation.

Grammar Items List In this study, we use the CEFR-J Grammar Profile (Ishii and Tono, 2018) as the grammar item list for evaluating GEC models.[4] The CEFR-J Grammar Profile primarily targets grammar items taught to English learners in Japan and incorporates insights from other established frameworks, including the T-series (Van Ek and Trim, 1991, 2001a,b), the Core Inventory for General English[5], and the English Grammar Profile[6]. It defines 263 grammar items, each mapped to a specific CEFR level (see Table 2 for examples). The reason we chose the CEFR-J Grammar Profile is that it defines grammar items not only based on observations of learner errors but also on the learning objectives specified for each CEFR level. This dual foundation makes it particularly well suited for evaluating GEC models at specific proficiency levels, as it allows us to assess whether a model can handle grammar items that learners are expected to master. Furthermore, because the grammar items closely reflect what annotators are accustomed to teaching, they support the creation of authentic and pedagogically meaningful minimal pairs.

[4]CEFR-J is a framework adapted from CEFR for English education in Japan. As part of the CEFR-J project, the CEFR-J Grammar Profile was developed to provide a grammar item list aligned with this framework. For this study, we used version 20200220 of the CEFR-J Grammar Profile.

[5]https://www.eaquals.org/resources/the-core-inventory-for-general-english

[6]https://www.englishprofile.org/english-grammar-profile

3.2 Construction Scheme

This subsection outlines the procedures and guidelines for dataset creation.

Procedures Each annotator described in Section 3.1 followed the steps below to create the dataset. Since GEC corrections can have multiple valid forms (Bryant and Ng, 2015; Choshen and Abend, 2018), the process involved generating both minimal pairs of ungrammatical and grammatical sentences as well as multiple grammatical variants for a single ungrammatical sentence.

Step 1. Identify and understand the target grammar item.

Step 2. Construct an ungrammatical sentence containing an error related to the target grammar item, along with its corresponding grammatical version that correctly applies the item.

Step 3. If alternative corrections exist, generate additional grammatical sentences that do not depend on the target grammar item.

Each annotator followed this procedure for every grammar item, generating three sentence pairs per item.[7] During Step 1, if ambiguities arose regarding the target grammar item, annotators and authors discussed them for clarification. In contrast, for Steps 2 and 3, each annotator worked independently to create ungrammatical and grammatical sentences. To ensure quality, we hired an additional

[7]The number of sentence pairs for each grammar item was determined based on budgetary constraints.

checker[8], in addition to the two annotators, to assess the produced text for compliance with the rules described below and to provide feedback. Sentence pairs deemed inappropriate by the checker were returned to the annotators for revision, and this process was repeated until the checker approved the sentence pair.[9]

Rules To ensure the reliability of the dataset, we established four guidelines for sentence creation.

Rule 1. Errors in ungrammatical sentences must be limited to the target grammar item.

Rule 2. In Step 3, grammatical sentences should be generated for all relevant correction patterns.

Rule 3. Sentences should be simple and reflect the writing style typically used by language learners.

Rule 4. Within each grammar item, errors should reflect common mistakes made by English learners.

Since GEC errors can interact, Rule 1 prevents unintended errors from affecting correction performance (Mita and Yanaka, 2021). Reference-based evaluation methods may penalize valid corrections if they do not appear in the reference sentence (Bryant and Ng, 2015; Choshen and Abend, 2018). Rule 2 ensures a more comprehensive assessment of GEC model performance. Additionally, since GEC models are influenced by the text domain (Napoles et al., 2019; Mita et al., 2019; Flachs et al., 2020), Rules 3 and 4 help ensure that the evaluation conditions closely resemble real-world applications.[10] Table 2 presents examples of actual minimal pairs across different CEFR levels. Each ungrammatical sentence contains only one error related to the target grammar item. The grammatical sentence in Step 2 corrects the error using the target grammar item, while the sentences in Step 3 employ alternative correction strategies. Alternative corrections in Step 3 were generated only when applicable.

[8]The reviewer is a certified English teacher with approximately 20 years of experience in developing educational materials for middle and high school English language learners.

[9]Approximately 29% of the sentence pairs were returned to the annotators by the checker at least once for revision.

[10]In this study, we developed CTSEG following the conventions of American English typically taught to English learners.

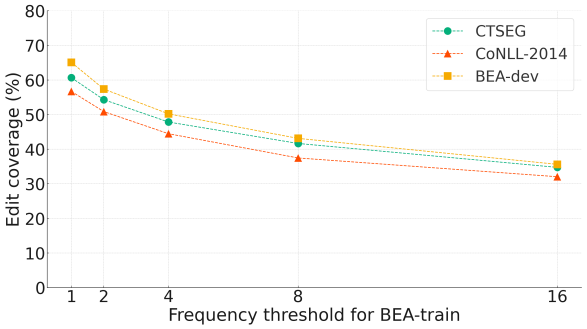


Figure 2: Edit coverage of each test/dev set relative to BEA-train, computed by varying the frequency threshold and retaining only edits that appear in BEA-train at or above that threshold.

Post-processing To prepare the dataset for GEC model evaluation, we applied the following post-processing steps. First, we tokenized all sentences using the spaCy toolkit (Honnibal and Montani, 2017). Second, we converted the dataset into M^2 format using ERRANT.[11] Finally, we divided the resulting M^2 file into separate files for each grammar item. These M^2 files will be publicly available for research purposes.

3.3 Edit Coverage

In this study, experienced English teachers created erroneous sentences to mimic common learner errors, which were then carefully reviewed by an independent checker. This section investigates the extent to which such artificially constructed errors overlap with those found in actual learner corpora. Specifically, we analyze how many of the edits in CTSEG overlap with those in BEA-train—a learner corpus consisting of FCE, the Lang-8 Corpus (Mizumoto et al., 2011), NUCLE, and W&I+LOCNESS, containing 564,684 sentence pairs. The overlap between CTSEG and BEA-train, referred to as edit coverage, is calculated as follows:

$$\text{Edit coverage} := \frac{|E_{\text{CTSEG}} \cap E_{\text{BEA-train}}|}{|E_{\text{CTSEG}}|} \quad (1)$$

Here, E_{CTSEG} denotes the set of edits in CTSEG, and $E_{\text{BEA-train}}$ denotes the set of edits in BEA-train.

Figure 2 presents the edit coverage for CTSEG. For comparison, we also measured the edit coverage of CoNLL-2014 and BEA-dev using the same

[11]We use ERRANT only to convert the data into M^2 format for evaluation with the M^2 scorer (Dahlmeier and Ng, 2012). In the experiments, we evaluate each individual M^2 file based on CTSEG error categories rather than ERRANT labels.

method. To focus on more frequently occurring errors, we applied a frequency threshold to the edits in BEA-train. A threshold of 1 includes all edits, while a threshold of 2 includes only those that appear at least twice. As shown in Figure 2, approximately 60% of the edits in CTSEG also appear at least once in BEA-train. This ratio is comparable to those of CoNLL-2014 and BEA-dev, indicating that CTSEG effectively mimics learner-like errors. Furthermore, even as the frequency threshold increases, the edit coverage of CTSEG remains similar to that of CoNLL-2014 and BEA-dev. This suggests that CTSEG captures frequent edits typical of actual learner corpora.

3.4 Comparing Datasets

Table 1 summarizes the dataset statistics, including its size, the number of reference sentences, the number of error tag types, and CEFR levels. Our dataset contains a comparable number of sentences to the CoNLL-2014 shared task test set, a widely used GEC benchmark. Moreover, each ungrammatical sentence is assigned up to three grammatical reference sentences when necessary. A notable distinction of our dataset is its large number of error tags, totaling 263, which reflects its foundation in the CEFR-J Grammar Profile. Additionally, our dataset covers grammar items from CEFR levels A1 to B2, aligning with CEFR-J. However, it does not yet include C1 or C2-level grammar items, highlighting the need for future expansion.

To compare correction tendencies, we analyze the distribution of edit distances (Levenshtein, 1966) across datasets. Figure 3 shows word-level edit distance distributions for CTSEG, the CoNLL-2014 test set, and the JFLEG test set.[12] We use the first target sentence in each dataset as the reference. CoNLL-2014 follows a minimal edit approach, while JFLEG focuses on fluency edits. Figure 3 shows that CTSEG has a smaller edit distance than CoNLL-2014, as it targets one error per sentence, allowing for item-specific evaluation. Since real-world sentences often contain multiple errors, existing benchmarks better reflect practical GEC conditions. Thus, depending on the evaluation goal, CTSEG and existing benchmarks can be used separately or together.

[12]Appendix B shows example sentence pairs from CTSEG for each word-level edit distance.

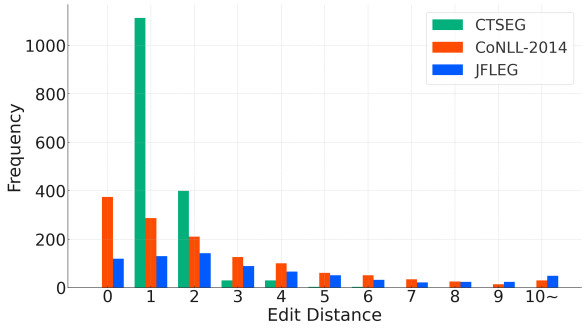


Figure 3: Distribution of word-level edit distances between source and target sentences across datasets.

4 Experiments

4.1 GEC Models

As mentioned in Section 2.2, current GEC research is primarily guided by three main paradigms. For our experiments, we selected representative models from each paradigm, as detailed below.

Transformer This model follows an encoder-decoder architecture (Vaswani et al., 2017). We used the Transformer-based model proposed by Kiyono et al. (2020), pre-trained on 70 million pseudo data and fine-tuned on real data.[13] Previous studies (Kiyono et al., 2020; Kaneko et al., 2020) have shown that this paradigm generally achieves higher recall than sequence tagging models like GECToR (Omelianchuk et al., 2020).

GECToR This model is based on a sequence tagging architecture. We adopted the model proposed by Omelianchuk et al. (2020), which employs approximately 5,000 tags, and RoBERTa (Liu et al., 2019) for tag prediction.[14] Prior research (Omelianchuk et al., 2020; Tarnavskyi et al., 2022) has demonstrated that sequence tagging models generally achieve higher precision than encoder-decoder architectures.

GPT-4 This model is a large language model (LLM) that generates corrections based on prompts. We used GPT-4 (OpenAI et al., 2024)[15], developed by OpenAI[16]. A key characteristic of GPT-4 is that its output varies based on prompt design (Coyne et al., 2023; Loem et al., 2023; Davis et al., 2024). For this study, we employed the English teacher

[13]https://github.com/butsugiri/gec-pseudodata
[14]https://github.com/grammarly/gector
[15]gpt-4-0613
[16]https://openai.com

prompt from Davis et al. (2024), which achieved the highest performance in their study.[17]

4.2 Evaluation

We evaluate model performance using our dataset, considering two types of scores: *overall scores* and *individual scores*. Overall scores aggregate grammar item scores across the dataset or by CEFR level, providing insights into general correction trends in GEC models. In contrast, individual scores assess performance for each grammar item separately, enabling a fine-grained analysis of a model's ability to correct specific grammatical errors. The choice between overall and individual scores depends on the evaluation objective.

The core contribution of our evaluation paradigm lies in **individual scores**, which allow for a detailed analysis of model performance across specific grammatical items. This level of granularity is essential for diagnosing strengths and weaknesses that remain hidden in conventional aggregate evaluations. Since our evaluation paradigm is designed to **complement rather than replace** existing leaderboard-style evaluations, our primary focus is on assessing how accurately models correct specific grammatical items. While conventional evaluations offer a broad assessment of overall model capability, our approach enables a more targeted analysis by examining correction accuracy at the individual grammar item level.

To compute these scores, we use the M^2 scorer (Dahlmeier and Ng, 2012). For overall evaluation, we report Precision, Recall, and $F_{0.5}$, whereas for individual evaluation, we focus on Recall, as it directly measures the extent to which the targeted errors are corrected. By integrating both overall and individual evaluations, our framework provides a comprehensive perspective on GEC model performance, balancing general correction trends with fine-grained linguistic insights.

4.3 Results

Overall Scores Table 3 summarizes model performance across the entire dataset and by CEFR level. GPT-4 outperforms both Transformer and GECToR across all metrics. Consistent with previous studies (Omelianchuk et al., 2020; Tarnavskyi et al., 2022), Transformer achieves relatively high recall, reflecting its strength in covering a broad range of errors. GECToR demonstrates higher precision, which aligns with its sequence tagging-based architecture. Regarding CEFR-level performance, all models exhibit notably lower performance at the B2 level compared to other levels. This suggests that complex grammar items at this level pose greater challenges for GEC models. While overall scores reveal general trends, the following analysis of individual scores focuses on specific grammar items where performance is particularly weak.

GEC Model	Category	P	R	$F_{0.5}$
Transformer	ALL	76.81	77.25	76.53
	A1	78.91	79.52	78.74
	A2	79.41	81.75	79.58
	B1	80.42	79.25	79.79
	B2	69.63	70.34	69.35
GECToR	ALL	79.83	57.83	71.66
	A1	85.11	60.42	75.07
	A2	84.80	63.80	77.79
	B1	78.42	57.44	70.42
	B2	73.75	52.19	66.15
GPT-4	ALL	83.47	85.29	83.64
	A1	85.68	86.48	85.68
	A2	85.05	86.81	85.22
	B1	86.36	87.78	86.49
	B2	77.64	80.74	78.00

Table 3: Overall scores for each GEC model. The values represent macro-average scores for grammar items in each category. Darker shades of red indicate lower scores.

Individual Scores Table 4 presents the five grammar items with the lowest average recall for each CEFR level. The results demonstrate that GEC model performance varies across different grammar items. For example, for the grammar item "Imperative sentences (general verb)," GPT-4 achieves a recall of 83.33, whereas GECToR's recall is 0. Beyond this specific grammar item, GECToR also fails to correct errors in other cases, with 8 out of 263 grammar items receiving no corrections at all. This limitation could not be identified using conventional benchmarks, highlighting the advantage of our proposed CTSEG benchmark.

Table 4 also indicates that certain grammar items, such as questions and imperative sentences, are hard to correct across the representative models. These grammar items are rarely used in existing benchmarks, such as the CoNLL-2014 test set, due to the nature of their content. As a result, despite their low performance, these grammar items have been largely underlooked in existing benchmarks. To enhance the effectiveness of GEC models in sup-

[17] For specific prompts, see the first row of Table 12 in Appendix C.

ID	CL	Grammar Item	Transformer	GECToR	GPT-4	Avg.
168	A1	Present participle modifying a noun (pre-position)	50.00	0.00	66.67	38.89
241	A1	Interrogative sentence (Which . . .?)	50.00	33.33	50.00	44.44
117	A1	Imperative sentences (general verb)	66.67	0.00	83.33	50.00
263	A1	Functional interrogative (What about . . .?)	57.14	42.86	57.14	52.38
25	A1	Indefinite pronoun (supportive word "ones")	33.33	66.67	66.67	55.56
244	A2	Interrogative sentence (Whose + noun . . .?)	16.67	16.67	66.67	33.33
243	A2	Interrogative sentence (Whose . . .?)	30.00	10.00	60.00	33.33
191	A2	Exclamatory sentence (How alone)	50.00	16.67	50.00	38.89
118	A2	Emphasis in affirmative imperative sentences with do	50.00	50.00	50.00	50.00
179	A2	Omission of relative pronoun (objective)	50.00	50.00	66.67	55.56
30	B1	Pronoun (the other/others)	16.67	33.33	50.00	33.33
116	B1	Imperative sentences (be verb)	33.33	0.00	66.67	33.33
45	B1	so + adjective/adverb + (that) clause	57.14	0.00	57.14	38.10
220	B1	as if/though + subjunctive past	50.00	16.67	50.00	38.89
184	B1	Preposition + relative pronoun	66.67	16.67	50.00	44.44
214	B2	Participial construction (past participle at the beginning)	16.67	0.00	33.33	16.67
182	B2	Pseudo-relative pronoun (as)	28.57	0.00	50.00	26.19
213	B2	Participial construction (present participle at the beginning)	42.86	14.29	28.57	28.57
80	B2	Passive voice (future continuous)	41.67	16.67	41.67	33.33
218	B2	wish + subjunctive past	33.33	33.33	33.33	33.33

Table 4: Recall for each grammar item across GEC models. For each CEFR level, the five grammar items with the lowest average recall were selected.

porting language learners, future research should focus on improving the performance of these underrepresented grammar items.

5 Leveraging GPT-4 to Support Language Learners

In the previous section, we found that the performance of GEC models is low for certain grammar items. In this section, we examine whether incorporating grammar item information into prompts enhances GPT-4's performance, assuming that it will be used to assist language learners. Additionally, we investigate whether specifying grammar items enables GPT-4 to refine its corrections to ensure the inclusion of the target grammar items.

5.1 Motivation

An example application of GEC models is supporting learners in verifying a specific grammar item they are striving to master. For instance, after learning relative pronouns in class, a learner writes a sentence using a relative pronoun and inputs it into a GEC model to check whether it is used correctly. In such cases, it is reasonable to assume that information about the target grammar item can be integrated into the prompts. Furthermore, from the perspective of learner support, the model should not only correct errors but also ensure that the target grammar item is preserved in the correction. However, since GEC models often allow for multiple valid corrections, the learner's intended grammar item may not always be present in the output.

To address this issue, we investigate whether modifying the prompt enhances GPT-4's performance and enables it to generate corrections that explicitly include the target grammar item. This experiment assumes a scenario in which the learner's target grammar item is predetermined, as described in the use case above. Therefore, we hypothesize that explicitly incorporating information about the target grammar item into the prompt can effectively guide GPT-4's output.

5.2 Settings

Prompts We evaluate two types of prompts: a baseline prompt and a grammar-item-guided prompt. The baseline prompt, adapted from Davis et al. (2024) and used in Section 4, allows GPT-4 to generate corrections without imposing any constraints on grammar item usage. In contrast, the grammar-item-guided prompt explicitly specifies the target grammar item that the learner aims to practice.[18] By explicitly guiding GPT-4, this prompt encourages the model to prioritize corrections that incorporate the specified grammar item.

Evaluation In addition to standard metrics, we introduce a novel metric called *targeted recall* to

[18]For specific prompts, see the second row of Table 12 in Appendix C.

Sentence (target grammar item: "Participial construction (present participle at the beginning)")		Perplexity
Source	**The boy seeing** me , he ran away .	289.48
Baseline prompt (same as a reference at Step 3)	**When the boy saw** me, he ran away .	23.59
Grammar-item-guided prompt (same as a reference at Step 2)	**Seeing** me , the boy ran away .	42.50

Table 5: Example outputs for each prompt with the target grammar item.

evaluate whether GEC models' corrections incorporate the target grammar item. Specifically, we assess the model using only the grammatical sentences generated in Step 2 of the dataset construction procedure as reference sentences. Since Step 2 involves generating grammatical sentences that explicitly include the target grammar item, this evaluation allows us to quantify the extent to which the model adheres to the target grammar item constraint in its corrections.

5.3 Results

Table 6 presents the evaluation results for each prompt. Single Ref refers to the evaluation using only grammatical sentences that include the target grammar item, while All Ref refers to the evaluation using all grammatical sentences. As shown in Table 6, GPT-4's performance improves when incorporating the target grammar item into the prompt. The baseline prompt achieves a recall of 78.33 when evaluated against grammar-item-specific references, whereas explicitly specifying the target grammar item in the prompt increases recall to 91.67.

Table 5 provides output examples from each prompt along with sentence perplexity scores. With the baseline prompt, GPT-4 generates fluent corrections but does not necessarily include the target grammar item. In contrast, when the target grammar item is explicitly specified in the prompt, GPT-4 produces corrections incorporating the specified grammar item, albeit with slightly higher perplexity. This outcome aligns with language learners' needs, as it ensures that their intended grammar practice is reinforced in the corrected output. These results demonstrate that prompt engineering is an effective approach for controlling GPT-4's output in grammar-focused learning applications.

6 Conclusions

In this study, we proposed integrating TSE into the GEC evaluation framework. As the first evaluation dataset based on TSE, we developed CTSEG to facilitate proficiency-aligned evaluations of GEC models. CTSEG is a relatively small dataset, with a size comparable to that of CoNLL-2014 and JFLEG. Nevertheless, as shown in Table 4, CTSEG successfully reveals model-specific weaknesses that are not easily captured by existing benchmarks, even with its limited size. These findings highlight the distinct capabilities and limitations of different models, thereby offering valuable insights for future development and evaluation of GEC systems.

Prompt	Single Ref	All Ref
Baseline	78.33	85.29
Grammar-item-guided	91.67	94.02
Δ	+13.34	+8.73

Table 6: Targeted recall of GPT-4 for each prompt.

We hope that CTSEG will be used alongside existing benchmarks. CTSEG enables fine-grained analysis of model weaknesses, making it a valuable resource for comprehensive GEC evaluation. For future work, given that manual dataset creation is resource-intensive, we plan to explore automated generation methods—including rule-based approaches and data augmentation using large language models—to enhance scalability. In addition, we aim to investigate the extent to which humans can correct errors made by annotators in CTSEG. This investigation will help determine whether the errors that GEC models fail to correct can instead be corrected by humans, or whether they are inherently difficult for both models and humans to address. We believe that this analysis will provide valuable insights for advancing GEC research.

Limitations

Dataset Size Limitation One limitation of this study is the relatively small number of minimal pairs per grammar item. In this study, minimal pairs were manually created by two English teachers to ensure high-quality evaluation data that adhered to the rules outlined in Section 3.2. To construct larger and more reliable evaluation datasets in the future, it is essential to explore methods for automating the generation of minimal pairs.

Error Operation Ratio Limitation One limitation of CTSEG is that it does not control the ratio of error operations (i.e., Missing, Replacement, Unnecessary). As a result, the distribution of error operations may differ from that of existing test sets, such as CoNLL-2014 and BEA-2019. In this study, we deliberately chose not to control the operation ratio in order to reduce the burden on annotators and to allow for the creation of more natural, learner-like errors. However, Davis et al. (2024) has shown that GEC model performance can vary depending on the type of error operation, suggesting that evaluation at the operation level is important for GEC research. To enable more fine-grained analysis, it would be desirable in future work to construct test sets in which the distribution of error operations is explicitly controlled for each grammar item.

Language Limitation In this study, we constructed a dataset for English to facilitate the evaluation of GEC models at the grammar item level. However, GEC research has also been conducted in languages such as Chinese (Tang et al., 2023; Yang and Quan, 2024) and Russian (Rozovskaya and Roth, 2019; Palma Gomez and Rozovskaya, 2024), where learner-produced texts are utilized in a similar manner to English. Therefore, grammar-item-specific evaluation datasets can also be developed for languages other than English.

Ethical Considerations

We collaborated with an annotation company to create our evaluation dataset. Through this company, we provided appropriate compensation to two English teachers and a reviewer. Specifically, each English teacher received 300 yen per sentence, while the reviewer was compensated 50 yen per sentence. Including fees and commissions, the total cost of dataset production amounted to approximately 1.5 million yen, exceeding the average wage in Japan.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Numbers 22H03651 and 25K03178.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 Shared Task on Grammatical Error Correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 793–805.

Christopher Bryant and Hwee Tou Ng. 2015. How Far are We from Fully Automatic High Quality Grammatical Error Correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 697–707.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical Error Correction: A Survey of the State of the Art. *Computational Linguistics*, pages 643–701.

Leshem Choshen and Omri Abend. 2018. Inherent Biases in Reference-based Evaluation for Grammatical Error Correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 632–642.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the Performance of GPT-3.5 and GPT-4 in Grammatical Error Correction. *Preprint*, arXiv:2303.14342.

Daniel Dahlmeier and Hwee Tou Ng. 2011. Grammatical Error Correction with Alternating Structure Optimization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 915–923.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better Evaluation for Grammatical Error Correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31.

Vidas Daudaravicius, Rafael E. Banchs, Elena Volodina, and Courtney Napoles. 2016. A Report on the Automatic Evaluation of Scientific Writing Shared Task. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 53–62.

Christopher Davis, Andrew Caines, O Andersen, Shiva Taslimipoor, Helen Yannakoudakis, Zheng Yuan, Christopher Bryant, Marek Rei, and Paula Buttery. 2024. Prompting open-source and commercial language models for grammatical error correction of English learner text. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11952–11967.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is ChatGPT a Highly Fluent Grammatical Error Correction System? A Comprehensive Evaluation. *Preprint*, arXiv:2304.01746.

Mariano Felice, Christopher Bryant, and Ted Briscoe. 2016. Automatic Extraction of Learner Errors in ESL Sentences Using Linguistically Enhanced Alignments. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 825–835.

Simon Flachs, Ophélie Lacroix, Helen Yannakoudakis, Marek Rei, and Anders Søgaard. 2020. Grammatical Error Correction in Low Error Density Domains: A New Benchmark and Analyses. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8467–8478.

Michael Gamon, Jianfeng Gao, Chris Brockett, Alexandre Klementiev, William B. Dolan, Dmitriy Belenko, and Lucy Vanderwende. 2008. Using Contextual Speller Techniques and Language Modeling for ESL Error Correction. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 449–456.

Sylviane Granger. 1998. The computerized learner corpus: a versatile new source of data for SLA research. In *Learner English on Computer*, pages 3–18. Addison Wesley Longman.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting Grammaticality on an Ordinal Scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 174–180.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Yasutake Ishii and Yukio Tono. 2018. Investigating Japanese EFL Learners' Overuse/Underuse of English Grammar Categories and Their Relevance to CEFR Levels. In *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference*, pages 160–165.

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-Decoder Models Can Benefit from Pre-trained Masked Language Models in Grammatical Error Correction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4248–4254.

Shun Kiyono, Jun Suzuki, Tomoya Mizumoto, and Kentaro Inui. 2020. Massive Exploration of Pseudo Data for Grammatical Error Correction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2134–2145.

Vladimir Iosifovich Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring Effectiveness of GPT-3 in Grammatical Error Correction: A Study on Performance and Controllability in Prompt-Based Methods. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 205–219.

Rebecca Marvin and Tal Linzen. 2018. Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Masato Mita, Tomoya Mizumoto, Masahiro Kaneko, Ryo Nagata, and Kentaro Inui. 2019. Cross-Corpora Evaluation and Analysis of Grammatical Error Correction Models — Is Single-Corpus Evaluation Enough? In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1309–1314.

Masato Mita and Hitomi Yanaka. 2021. Do Grammatical Error Correction Models Realize Grammatical Generalization? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4554–4561.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining Revision Log

of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155.

Ryo Nagata, Edward Whittaker, and Vera Sheinman. 2011. Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1210–1219.

Courtney Napoles, Maria Nădejde, and Joel Tetreault. 2019. Enabling Robust Grammatical Error Correction in New Domains: Data Sets, Metrics, and Analyses. *Transactions of the Association for Computational Linguistics*, 7:551–566.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. 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*, pages 229–234.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – Grammatical Error Correction: Tag, Not Rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Frank Palma Gomez and Alla Rozovskaya. 2024. Multi-

Reference Benchmarks for Russian Grammatical Error Correction. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1253–1270.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A Simple Recipe for Multilingual Grammatical Error Correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 702–707.

Alla Rozovskaya and Dan Roth. 2019. Grammar Error Correction in Morphologically Rich Languages: The Case of Russian. *Transactions of the Association for Computational Linguistics*, 7:1–17.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the Goals of Grammatical Error Correction: Fluency Instead of Grammaticality. *Transactions of the Association for Computational Linguistics*, 4:169–182.

Taiga Someya, Ryo Yoshida, and Yohei Oseki. 2024. Targeted Syntactic Evaluation on the Chomsky Hierarchy. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 15595–15605.

Chenming Tang, Xiuyu Wu, and Yunfang Wu. 2023. Are Pre-trained Language Models Useful for Model Ensemble in Chinese Grammatical Error Correction? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 893–901.

Maksym Tarnavskyi, Artem Chernodub, and Kostiantyn Omelianchuk. 2022. Ensembling and Knowledge Distilling of Large Sequence Taggers for Grammatical Error Correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3842–3852.

J. A. Van Ek and J. L. M. Trim. 1991. *Threshold 1990*. Cambridge University Press.

J. A. Van Ek and J. L. M. Trim. 2001a. *Vantage*. Cambridge University Press.

J. A. Van Ek and J. L. M. Trim. 2001b. *Waystage 1990*. Cambridge University Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. ChatGPT or Grammarly? Evaluating ChatGPT on Grammatical Error Correction Benchmark. *Preprint*, arXiv:2303.13648.

Haihui Yang and Xiaojun Quan. 2024. Alirector: Alignment-Enhanced Chinese Grammatical Error Corrector. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2531–2546.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189.

Helen Yannakoudakis, Øistein E Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267.

Appendix

A CTSEG Error Categories and GEC Model Performance

The CTSEG error categories and the recall scores of each GEC model are presented in Tables 7–10. Compared to ERRANT, CTSEG provides a significantly more fine-grained classification of grammatical errors. For example, ERRANT's broad categories—such as VERB:TENSE or DET (determiners)—group together a variety of learner errors. In contrast, CTSEG, based on the CEFR-J Grammar Profile, breaks these down into more specific constructions. ERRANT's VERB:TENSE category, for instance, may map to several CTSEG items such as "present," "past," or "future." Likewise, DET may correspond to more specific categories in CTSEG like "a/an/the," "some/any," and "no."

However, it is difficult to fully map the CEFR-J-based error categories in CTSEG to ERRANT's error types due to fundamental differences in their design goals. CTSEG is pedagogically motivated and structured around what learners are expected to acquire at each proficiency level, whereas ERRANT is error-oriented and curriculum-agnostic. As a result, many CTSEG categories lack clear counterparts in ERRANT, making one-to-one mapping between the two frameworks challenging.

B Example Sentence Pairs by Edit Distance

Table 11 presents example sentence pairs corresponding to each word-level edit distance. Because

CTSEG instructs annotators to create minimal pairs, the edit distance is often between 1 and 2, as shown in Figure 3. However, depending on the grammar item, it can range from 3 to 5, as also illustrated in Table 11.

C Prompt for GPT-4

The prompts used for GPT-4 in the experiment are shown in Table 12.

ID	CL	Grammar Item	Transformer	GECToR	GPT-4	Avg.
1	A1	Personal pronoun (nominative case) (I) + be: I am	83.33	66.67	100.00	83.33
2	A1	Personal pronoun (nominative case) (you) + be: You are	100.00	100.00	100.00	100.00
3	A1	Personal pronoun (nominative case) (he/she) + be: He/She is	66.67	83.33	100.00	83.33
4	A1	Personal pronoun (nominative case) (we) + be: We are	100.00	83.33	100.00	94.44
5	A1	Personal pronoun (nominative case) (they) + be: They are	83.33	50.00	83.33	72.22
6	A1	Personal pronoun (possessive case): my/our/your/her/their	71.43	71.43	100.00	80.95
7	A1	Personal pronoun (objective case): me/us/him/her/them	100.00	66.67	83.33	83.33
8	A1	Demonstrative pronoun (this/that) + be: This/That is	75.00	75.00	100.00	83.33
9	A1	Demonstrative pronoun (these/those) + be: These/Those are	85.71	57.14	85.71	76.19
10	A1	Demonstrative pronoun (it) + be: It is	87.50	28.57	62.50	59.52
11	A1	Demonstrative adjective (this/that + noun)	62.50	28.57	100.00	63.69
12	A1	Demonstrative adjective (these/those + noun)	100.00	62.50	100.00	87.50
13	A1	Indefinite article	100.00	66.67	83.33	83.33
14	A1	Definite article	83.33	33.33	66.67	61.11
15	A1	Determiner (some/any)	83.33	50.00	100.00	77.78
16	A1	Determiner (no)	90.00	37.50	100.00	75.83
17	A1	Determiner (another)	85.71	50.00	100.00	78.57
18	A1	much + uncountable noun	100.00	42.86	100.00	80.95
19	A2	little + uncountable noun	83.33	33.33	83.33	66.67
20	A2	few + plural noun	83.33	66.67	85.71	78.57
21	A2	Prepositions	100.00	83.33	100.00	94.44
22	A1	Possessive pronouns	100.00	66.67	100.00	88.89
23	A2	Reflexive pronouns	62.50	37.50	87.50	62.50
24	A1	Indefinite pronouns (-thing/-one/-body)	71.43	42.86	85.71	66.67
25	A1	Indefinite pronoun (supportive word "ones")	33.33	66.67	66.67	55.56
26	A2	Indefinite pronoun (none)	87.50	25.00	100.00	70.83
27	B1	Reciprocal pronoun (each other)	100.00	57.14	83.33	80.16
28	B1	Reciprocal pronoun (one another)	75.00	42.86	77.78	65.21
29	B1	Pronoun (others) (excluding "the others")	100.00	50.00	83.33	77.78
30	B1	Pronoun (the other/others)	16.67	33.33	50.00	33.33
31	A2	-thing + adjective	100.00	66.67	100.00	88.89
32	A2	Adverb (frequency)	57.14	42.86	71.43	57.14
33	B1	Adverb (emphasis)	100.00	37.50	100.00	79.17
34	B2	Adverb (manner)	42.86	28.57	71.43	47.62
35	B1	Adverb (negation)	100.00	44.44	100.00	81.48
36	B1	Adverb (quasi-negation)	66.67	33.33	88.89	62.96
37	A2	Comparison of equality (as + adjective/adverb + as)	83.33	50.00	100.00	77.78
38	A1	Comparative (superiority) (-er) (including irregular forms like "better")	88.89	55.56	88.89	77.78
39	A2	Comparative (superiority) (more + adjective/adverb)	100.00	83.33	100.00	94.44
40	A1	Superlative (superiority) (-est) (including irregular forms like "best")	100.00	57.14	100.00	85.71
41	A2	Superlative (superiority) (most + adjective/adverb)	100.00	66.67	83.33	83.33
42	B1	Comparative (inferiority)	66.67	37.50	83.33	62.50
43	B1	Adjective/adverb + enough	100.00	28.57	100.00	76.19
44	B1	too + adjective/adverb + to-infinitive	100.00	66.67	100.00	88.89
45	B1	so + adjective/adverb + (that) clause	57.14	0.00	57.14	38.10
46	B1	such (+a/an) + adjective + noun	100.00	66.67	100.00	88.89
47	B1	so + adjective + a/an + noun	100.00	33.33	100.00	77.78
48	B1	too + adjective + a/an + noun	71.43	0.00	90.91	54.11
49	B1	Comparative and comparative (same comparative)	45.45	18.18	100.00	54.55
50	B2	the + comparative (...), the + comparative	88.89	44.44	88.89	74.07
51	B1	Emphasis of comparative (e.g., even)	85.71	55.56	83.33	74.87
52	B1	Emphasis of superlative (e.g., by far)	50.00	37.50	62.50	50.00
53	B1	Emphasis with do/does	33.33	50.00	83.33	55.56
54	B1	Emphasis with did	83.33	50.00	83.33	72.22
55	B2	Phrasal verbs (verb + particle)	85.71	14.29	100.00	66.67
56	B2	Phrasal verbs (verb + noun phrase/pronoun + particle)	83.33	50.00	83.33	72.22
57	B2	Phrasal verbs (verb + particle + preposition + noun phrase/pronoun)	83.33	50.00	100.00	77.78
58	A1	Tense/aspect (present) (be verb)	57.14	57.14	85.71	66.67
59	A1	Tense/aspect (present) (general verb, except third-person singular)	100.00	100.00	83.33	94.44
60	A1	Tense/aspect (present) (general verb, third-person singular)	100.00	57.14	100.00	85.71
61	A1	Tense/aspect (present continuous)	71.43	71.43	71.43	71.43
62	A2	Tense/aspect (present perfect)	66.67	50.00	100.00	72.22
63	B2	Tense/aspect (present perfect continuous)	100.00	66.67	100.00	88.89
64	A1	Tense/aspect (past) (be verb)	83.33	83.33	83.33	83.33
65	A1	Tense/aspect (past) (general verb)	100.00	100.00	100.00	100.00
66	A2	Tense/aspect (past continuous)	100.00	100.00	100.00	100.00
67	B1	Tense/aspect (past perfect)	66.67	66.67	83.33	72.22
68	B2	Tense/aspect (past perfect continuous)	50.00	16.67	66.67	44.44
69	A1	Tense/aspect (future)	66.67	33.33	66.67	55.56
70	B1	Tense/aspect (future continuous)	50.00	83.33	83.33	72.22
71	B2	Tense/aspect (future perfect)	50.00	83.33	100.00	77.78
72	B2	Tense/aspect (future perfect continuous)	33.33	50.00	83.33	55.56
73	A1	Passive voice (present)	100.00	100.00	100.00	100.00
74	B2	Passive voice (present continuous)	100.00	66.67	83.33	83.33
75	B2	Passive voice (present perfect)	83.33	100.00	100.00	94.44
76	A2	Passive voice (past)	83.33	50.00	90.00	74.44
77	B2	Passive voice (past continuous)	66.67	50.00	50.00	55.56
78	B2	Passive voice (past perfect)	33.33	33.33	100.00	55.56
79	B1	Passive voice (future)	100.00	100.00	100.00	100.00
80	B2	Passive voice (future continuous)	41.67	16.67	41.67	33.33

Table 7: Recall for each grammar item (ID1–ID80) across GEC models.

ID	CL	Grammar Item	Transformer	GECToR	GPT-4	Avg.
81	B2	Passive voice (future perfect)	66.67	66.67	83.33	72.22
82	B1	Passive voice (modal verbs)	100.00	85.71	100.00	95.24
83	B2	Passive voice (modal verbs + continuous)	50.00	40.00	66.67	52.22
84	B2	Passive voice (modal verbs + perfect)	100.00	100.00	100.00	100.00
85	B2	Passive voice (give/pass/send/show/teach/tell with indirect object as subject)	54.55	54.55	72.73	60.61
86	B2	Passive voice (give/pass/send/show/teach/tell with direct object as subject)	100.00	83.33	83.33	88.89
87	B1	get + past participle	100.00	83.33	100.00	94.44
88	A2	to-infinitive (to DO)	100.00	75.00	100.00	91.67
89	B1	Negative form of to-infinitive (not to DO)	83.33	83.33	83.33	83.33
90	B2	Perfect to-infinitive	83.33	66.67	50.00	66.67
91	B2	Passive to-infinitive	50.00	50.00	50.00	50.00
92	B2	Perfect passive to-infinitive	100.00	100.00	83.33	94.44
93	B1	to-infinitive with logical subject	71.43	37.50	60.00	56.31
94	B1	-thing + to-infinitive	100.00	100.00	100.00	100.00
95	B2	in order to DO	71.43	42.86	85.71	66.67
96	B2	in order not to DO	71.43	57.14	100.00	76.19
97	B2	so as to DO	83.33	85.71	100.00	89.68
98	B2	so as not to DO	85.71	28.57	85.71	66.67
99	B1	be to DO	85.71	33.33	57.14	58.73
100	B2	be about to DO	100.00	66.67	100.00	88.89
101	A2	Verb + to-infinitive	100.00	100.00	100.00	100.00
102	B1	Verb + not + to-infinitive	100.00	14.29	100.00	71.43
103	A1	Verb + object + to-infinitive	100.00	83.33	100.00	94.44
104	B2	Verb + object + not + to-infinitive	70.00	0.00	66.67	45.56
105	A2	Verb -ing form	83.33	66.67	83.33	77.78
106	B2	not + verb -ing form	50.00	16.67	66.67	44.44
107	B2	having + past participle	50.00	50.00	83.33	61.11
108	B2	being + past participle	50.00	66.67	100.00	72.22
109	B2	having been + past participle	33.33	66.67	66.67	55.56
110	A1	Preposition + verb -ing form	83.33	83.33	83.33	83.33
111	B2	Gerund with logical subject (possessive pronoun)	66.67	50.00	100.00	72.22
112	A2	Verb + verb -ing form	83.33	83.33	83.33	83.33
113	B2	Verb + not + verb -ing form	83.33	66.67	83.33	77.78
114	A2	Verb + object + verb -ing form	66.67	50.00	83.33	66.67
115	B2	Verb + object + not + verb -ing form	42.86	42.86	87.50	57.74
116	B1	Imperative sentences (be verb)	33.33	0.00	66.67	33.33
117	A1	Imperative sentences (general verb)	66.67	0.00	83.33	50.00
118	A2	Emphasis in affirmative imperative sentences with do	50.00	50.00	50.00	50.00
119	A1	Please + affirmative imperative sentence	83.33	85.71	55.56	74.87
120	A1	Let's	75.00	62.50	75.00	70.83
121	B1	Modal verbs (be able to)	100.00	66.67	100.00	88.89
122	A2	Modal verbs (be going to)	100.00	83.33	83.33	88.89
123	A2	Modal verbs (can)	100.00	85.71	100.00	95.24
124	B1	Modal verbs (could)	71.43	100.00	100.00	90.48
125	B2	Modal verbs (dare (to))	42.86	28.57	71.43	47.62
126	B1	Modal verbs (had better)	75.00	28.57	85.71	63.10
127	A2	Modal verbs (have to)	100.00	100.00	100.00	100.00
128	B1	Modal verbs ((have) got to)	57.14	71.43	85.71	71.43
129	B1	Modal verbs (may)	33.33	50.00	83.33	55.56
130	B2	Modal verbs (may as well)	50.00	33.33	66.67	50.00
131	B2	Modal verbs (may well)	66.67	66.67	83.33	72.22
132	B1	Modal verbs (might)	100.00	66.67	100.00	88.89
133	B1	Modal verbs (might as well)	100.00	50.00	100.00	83.33
134	B2	Modal verbs (might well)	77.78	66.67	66.67	70.37
135	B1	Modal verbs (must)	100.00	83.33	100.00	94.44
136	A2	Modal verbs (need (to))	85.71	57.14	71.43	71.43
137	B1	Modal verbs (ought to)	100.00	100.00	100.00	100.00
138	B1	Modal verbs (shall)	83.33	83.33	100.00	88.89
139	B1	Modal verbs (should)	100.00	100.00	100.00	100.00
140	B1	Modal verbs (used to)	83.33	66.67	83.33	77.78
141	A2	Modal verbs (will)	66.67	57.14	66.67	63.49
142	B1	Modal verbs (would)	100.00	57.14	100.00	85.71
143	B1	Modal verbs (would rather)	100.00	83.33	83.33	88.89
144	B1	Modal verbs + continuous	50.00	50.00	50.00	50.00
145	B2	Modal verbs + perfect	100.00	100.00	100.00	100.00
146	A1	There of existence (There + be)	100.00	50.00	100.00	83.33
147	B1	There of existence (perfect)	100.00	66.67	100.00	88.89
148	B1	There of existence (modal verbs)	57.14	42.86	71.43	57.14
149	A1	Here of existence (Here is/are)	57.14	57.14	57.14	57.14
150	A1	Coordinating conjunctions	71.43	66.67	71.43	69.84
151	B2	that-clause (appositive)	83.33	66.67	100.00	83.33
152	A1	that-clause (object)	85.71	71.43	85.71	80.95
153	B1	wh-clause (object) (excluding "whether")	100.00	85.71	85.71	90.48
154	B1	whether-clause	66.67	33.33	77.78	59.26
155	A1	Adverbial clause (when)	83.33	83.33	100.00	88.89
156	A2	Adverbial clause (if)	83.33	50.00	83.33	72.22
157	A2	Adverbial clause (as)	90.00	30.00	60.00	60.00
158	B1	Adverbial clause (as soon as)	100.00	25.00	100.00	75.00
159	B1	Adverbial clause (by the time)	100.00	100.00	100.00	100.00
160	B1	Adverbial clause (so that)	62.50	66.67	100.00	76.39

Table 8: Recall for each grammar item (ID81–ID160) across GEC models.

ID	CL	Grammar Item	Transformer	GECToR	GPT-4	Avg.
161	B1	Subordinate clauses (main subordinating conjunctions other than as/if/that/when/whether)	66.67	50.00	83.33	66.67
162	A1	Omission of that in subordinate clauses (e.g., hope/know/think)	66.67	66.67	83.33	72.22
163	A2	It as a formal subject + to-infinitive	85.71	57.14	100.00	80.95
164	B2	It as a formal object + to-infinitive	100.00	100.00	100.00	100.00
165	B1	It as a formal subject + that-clause	71.43	33.33	85.71	63.49
166	B1	It as a formal object + that-clause	60.00	77.78	85.71	74.50
167	B1	wh-word + to-infinitive	66.67	66.67	83.33	72.22
168	A1	Present participle modifying a noun (pre-position)	50.00	0.00	66.67	38.89
169	A1	Present participle modifying a noun (post-position)	75.00	83.33	100.00	86.11
170	A1	Past participle modifying a noun (pre-position)	57.14	33.33	85.71	58.73
171	A1	Past participle modifying a noun (post-position)	85.71	85.71	100.00	90.48
172	B1	Relative pronoun (nominative) (who)	100.00	83.33	100.00	94.44
173	B1	Relative pronoun (nominative) (which)	100.00	50.00	100.00	83.33
174	B1	Relative pronoun (nominative) (that)	50.00	33.33	50.00	44.44
175	B2	Relative pronoun (objective) (who)	100.00	83.33	100.00	94.44
176	B2	Relative pronoun (objective) (whom)	100.00	50.00	100.00	83.33
177	B2	Relative pronoun (objective) (which)	100.00	66.67	100.00	88.89
178	B2	Relative pronoun (objective) (that)	83.33	50.00	83.33	72.22
179	A2	Omission of relative pronoun (objective)	50.00	50.00	66.67	55.56
180	B1	Relative pronoun (possessive)	100.00	50.00	100.00	83.33
181	B1	Relative pronoun (non-restrictive clause)	83.33	66.67	83.33	77.78
182	B2	Pseudo-relative pronoun (as)	28.57	0.00	50.00	26.19
183	A2	Compound relative pronoun (what)	83.33	66.67	83.33	77.78
184	B1	Preposition + relative pronoun	66.67	16.67	50.00	44.44
185	A2	Relative adverb (with antecedent)	83.33	50.00	83.33	72.22
186	B1	Relative adverb (without antecedent)	83.33	66.67	83.33	77.78
187	B1	Relative adverb (non-restrictive clause)	66.67	66.67	83.33	72.22
188	B2	wh-ever	50.00	50.00	83.33	61.11
189	B1	Preposition stranding in relative clauses	71.43	71.43	81.82	74.89
190	A2	Exclamatory sentence (How + adjective/adverb)	100.00	42.86	85.71	76.19
191	A2	Exclamatory sentence (How alone)	50.00	16.67	50.00	38.89
192	A2	Exclamatory sentence (What)	71.43	71.43	71.43	71.43
193	B1	Tag question (following affirmative sentence)	57.14	28.57	100.00	61.90
194	A2	S + V	66.67	66.67	66.67	66.67
195	B1	S + V (become/feel/go/look/seem/sound) + C (adjective)	100.00	66.67	100.00	88.89
196	A2	S + V + O	100.00	85.71	100.00	95.24
197	B1	S + V (give/pass/send/show/teach/tell) + indirect object + direct object	100.00	87.50	100.00	95.83
198	B1	S + V (give/pass/send/show/teach/tell) + direct object + to + indirect object	100.00	71.43	100.00	90.48
199	B1	S + V (make) + O + C (adjective)	87.50	62.50	87.50	79.17
200	A2	Reported speech (explain/report/say)	83.33	100.00	100.00	94.44
201	B1	Reported speech (tell)	83.33	83.33	100.00	88.89
202	A2	Indirect questions (decide/explain/know/learn/see/understand/wonder)	85.71	71.43	85.71	80.95
203	B1	Indirect questions (ask/remind/show/teach/tell)	100.00	100.00	100.00	100.00
204	B2	Cleft sentence (emphasizing prepositional phrase/adverb)	50.00	16.67	62.50	43.06
205	B2	Pseudo-cleft sentence with what	66.67	33.33	83.33	61.11
206	A2	Causative construction (have/let/make)	100.00	83.33	100.00	94.44
207	B2	have/get + object + past participle	66.67	83.33	66.67	72.22
208	B1	get + object + present participle	66.67	100.00	100.00	88.89
209	B1	ask/tell + object + to-infinitive	83.33	66.67	83.33	77.78
210	B2	Perception verbs + bare infinitive	85.71	85.71	71.43	80.95
211	B2	Perception verbs + present participle	66.67	50.00	50.00	55.56
212	B2	Perception verbs + past participle	50.00	50.00	50.00	50.00
213	B2	Participial construction (present participle at the beginning)	42.86	14.29	28.57	28.57
214	B2	Participial construction (past participle at the beginning)	16.67	0.00	33.33	16.67
215	B2	Subjunctive past in if-clauses (past tense verb in if-clause)	50.00	33.33	100.00	61.11
216	B2	Subjunctive past perfect in if-clauses (past perfect verb in if-clause)	83.33	83.33	100.00	88.89
217	B2	Subjunctive present in that-clauses (bare infinitive in that-clause)	77.78	44.44	100.00	74.07
218	B2	wish + subjunctive past	33.33	33.33	33.33	33.33
219	B2	wish + subjunctive past perfect	83.33	66.67	100.00	83.33
220	B1	as if/though + subjunctive past	50.00	16.67	50.00	38.89
221	B2	as if/though + subjunctive past perfect	50.00	50.00	66.67	55.56
222	B2	if only + subjunctive past	83.33	50.00	83.33	72.22
223	B2	if only + subjunctive past perfect	100.00	66.67	66.67	77.78
224	B2	should in if-clauses	71.43	57.14	71.43	66.67
225	B2	Subjunctive inversion for past subjunctive	83.33	50.00	83.33	72.22
226	B2	Subjunctive inversion for past perfect subjunctive	85.71	33.33	100.00	73.02
227	B2	Inversion using should	83.33	66.67	100.00	83.33
228	B2	if it were/was not for ...	66.67	50.00	83.33	66.67
229	B2	were it not for ...	85.71	85.71	100.00	90.48
230	B2	if it hadn't been for ...	66.67	33.33	83.33	61.11
231	B2	had it not been for ...	100.00	50.00	100.00	83.33
232	B2	Inversion (So + be/have/do/modal verb + personal pronoun)	66.67	16.67	83.33	55.56
233	B2	Inversion (Neither/nor + be/have/do/modal verb + personal pronoun)	83.33	33.33	66.67	61.11
234	B2	Inversion (Hardly/Little/Never/No sooner/Scarcely/Seldom ...)	88.89	42.86	100.00	77.25
235	A1	Interrogative sentence (Why ...?)	62.50	75.00	77.78	71.76
236	A1	Interrogative sentence (When ...?)	88.89	55.56	88.89	77.78
237	A1	Interrogative sentence (Who ...?)	50.00	37.50	87.50	58.33
238	A1	Interrogative sentence (Whom ...?)	69.23	33.33	85.71	62.76
239	A1	Interrogative sentence (What ...?)	83.33	66.67	100.00	83.33
240	A1	Interrogative sentence (What + noun ...?)	62.50	50.00	62.50	58.33

Table 9: Recall for each grammar item (ID161–ID240) across GEC models.

ID	CL	Grammar Item	Transformer	GECToR	GPT-4	Avg.
241	A1	Interrogative sentence (Which ...?)	50.00	33.33	50.00	44.44
242	A1	Interrogative sentence (Which + noun ...?)	88.89	33.33	100.00	74.07
243	A2	Interrogative sentence (Whose ...?)	30.00	10.00	60.00	33.33
244	A2	Interrogative sentence (Whose + noun ...?)	16.67	16.67	66.67	33.33
245	A1	Interrogative sentence (Where ...?)	71.43	57.14	85.71	71.43
246	A2	Interrogative sentence (How ...?)	88.89	66.67	100.00	85.19
247	B1	Interrogative sentence (How + adjective/adverb ...?)	42.86	42.86	100.00	61.90
248	B1	Interrogative sentence (Preposition + interrogative ...?)	42.86	14.29	88.89	48.68
249	A1	Functional interrogative (Can you ...?)	100.00	66.67	100.00	88.89
250	A2	Functional interrogative (Could you ...?)	100.00	83.33	100.00	94.44
251	A2	Functional interrogative (Will you ...?)	100.00	85.71	100.00	95.24
252	A1	Functional interrogative (Would you ...?)	66.67	66.67	66.67	66.67
253	A1	Functional interrogative (Can I ...?)	66.67	83.33	100.00	83.33
254	A2	Functional interrogative (Could I ...?)	100.00	100.00	100.00	100.00
255	B1	Functional interrogative (May I ...?)	100.00	100.00	100.00	100.00
256	A2	Functional interrogative (Shall I ...?)	83.33	83.33	100.00	88.89
257	A2	Functional interrogative (Shall we ...?)	85.71	66.67	100.00	84.13
258	A2	Functional interrogative (Should I ...?)	100.00	83.33	100.00	94.44
259	A2	Functional interrogative (Why don't you ...?)	83.33	83.33	100.00	88.89
260	A2	Functional interrogative (Why don't we ...?)	66.67	55.56	83.33	68.52
261	A2	Functional interrogative (Why not ...?)	57.14	50.00	71.43	59.52
262	A1	Functional interrogative (How about ...?)	83.33	66.67	83.33	77.78
263	A1	Functional interrogative (What about ...?)	57.14	42.86	57.14	52.38

Table 10: Recall for each grammar item (ID241–ID263) across GEC models.

ID	Grammar Item	ED	Sentence
101	Verb + to-infinitive	1	*I want __ play soccer . I want **to** play soccer .
162	Omission of that in subordinate clauses (e.g., hope/know/think)	2	*We know **him** innocent . We know **he is** innocent .
11	Demonstrative adjective (this/that + noun)	3	*This **my dictionary** is good . This **dictionary of mine** is good .
198	S + V + direct object + to + indirect object (V: give/pass/send/show/teach/tell)	4	*I gave **to my friend the book** . I gave **the book to my friend** .
85	Passive voice (give/pass/send/show/teach/tell with indirect object as subject)	5	***A letter was sent my parents** by the teacher . **My parents were sent a letter** by the teacher .

Table 11: Examples of sentence pairs in CTSEG with word-level edit distances ranging from 1 to 5. The ED column indicates the word-level edit distance. An asterisk (*) denotes an ungrammatical sentence.

Name	Prompt
Baseline	You are an English language teacher. A student has sent you the following text. \n{text}\nProvide a grammatical correction for the text, making only necessary changes. Do not provide any additional comments or explanations. If the input text is already correct, return it unchanged.
Grammar-item-guided	You are an English language teacher. You taught {grammar_item} in today's class. A student has sent you the following text. \n{text}\nProvide a grammatical correction for the text, making only necessary changes. Do not provide any additional comments or explanations. If the input text is already correct, return it unchanged.

Table 12: Prompt for GPT-4: {text} represents the source sentence, and {grammar_item} specifies the target grammar item.