Is ChatGPT a Highly Fluent Grammatical Error Correction System? A Comprehensive Evaluation

Tao Fang[1]   **Shu Yang**[1]   **Kaixin Lan**[1]   **Derek F. Wong**[1*]   **Jinpeng Hu**[2]
Lidia S. Chao[1]   **Yue Zhang**[3]

[1]NLP[2]CT Lab, Department of Computer and Information Science, University of Macau
nlp2ct.{taofang,shuyang,kaixin}@gmail.com, {derekfw,lidiasc}@um.edu.mo
[2]The Chinese University of Hong Kong (Shenzhen)   [3]Soochow University
jinpenghu@link.cuhk.edu.cn   yzhang21@stu.suda.edu.cn

Abstract

ChatGPT, a large-scale language model based on the advanced GPT-3.5 architecture, has shown remarkable potential in various Natural Language Processing (NLP) tasks. However, there is currently a dearth of comprehensive study exploring its potential in the area of Grammatical Error Correction (GEC). To showcase its capabilities in GEC, we design zero-shot chain-of-thought (CoT) and few-shot CoT settings using in-context learning for ChatGPT. Our evaluation involves assessing ChatGPT's performance on five official test sets in three different languages, along with three document-level GEC test sets in English. Our experimental results and human evaluations demonstrate that ChatGPT has excellent error detection capabilities and can freely correct errors to make the corrected sentences very fluent, possibly due to its over-correction tendencies and not adhering to the principle of minimal edits. Additionally, its performance in non-English and low-resource settings highlights its potential in multilingual GEC tasks. However, further analysis of various types of errors at the document-level has shown that ChatGPT cannot effectively correct agreement, coreference, tense errors across sentences, and cross-sentence boundary errors.

1 Introduction

In recent years, Natural Language Processing (NLP) has made significant advancements due to the emergence of large language models (LLMs). Among the numerous LLMs available, Generative Per-trained Transformer (GPT) models (Radford et al., 2019; Brown et al., 2020) have demonstrated high efficacy in various NLP tasks. Recently, Chat-GPT[1], an advanced language model developed by OpenAI, has gained significant attention from researchers and practitioners in the field of NLP (Qin

---
*Corresponding Author
[1]https://openai.com/blog/chatgpt

et al., 2023; Luo et al., 2023; Liu et al., 2023; Hendy et al., 2023). Building upon the InstructGPT (Ouyang et al., 2022), ChatGPT is a groundbreaking innovation in the field of conversational agents. It possesses the remarkable ability to understand complex instructions and generate responses that closely mimic human speech. In addition to its conversational abilities, ChatGPT has demonstrated impressive performance in various other NLP tasks, including machine translation (Jiao et al., 2023; Hendy et al., 2023; Peng et al., 2023),question-answering (Bang et al., 2023), and text summarization (Yang et al., 2023).

When it comes to the Grammatical Error Correction (GEC) ability of ChatGPT, it appears that many individuals favor its use for text revision and refinement. Nonetheless, there is currently a scarcity of comprehensive research literature regarding ChatGPT's genuine error correction capabilities. To the best of our knowledge, only one study (Wu et al., 2023) conducts a preliminary evaluation of ChatGPT's performance in the English GEC task of the CoNLL14 test by only analyzing random 300 sentences and yields some preliminary but not sufficient results, which leaves us unclear about ChatGPT's specific capabilities and advantages in the GEC task. To address the current research gap, the aim of this study is to comprehensively investigate the performance and potential of ChatGPT in GEC, as well as to compare it with state-of-the-art (SOTA) models. This includes evaluating its performance at both the sentence-level and document-level GEC, designing prompts for zero-shot and few-shot scenarios, and assessing its effectiveness in English, non-English languages, and low-resource settings.

To explore the potential of ChaGPT for GEC, we first conduct a preliminary study to examine the effectiveness of our designed prompting methods in both **zero-shot** and **zero-shot chain-of-thought (CoT)** settings (Kojima et al., 2023). Our findings

indicate that incorporating CoT techniques can significantly enhance the performance of GEC on both CoNLL14 test and BEA19 test sets. Moreover, we propose the utilization of few-shot prompts in conjunction with CoT techniques to further enhance ChatGPT's performance through in-context learning (Brown et al., 2020). We evaluate ChatGPT's performance for sentence-level grammatical error correction (GEC) in three languages covered five official test sets using **zero-shot CoT**, **1-shot CoT**, **3-shot CoT**, and **5-shot CoT**. Besides, we also assess its performance in **zero-shot CoT** and **1-shot CoT** scenarios for document-level GEC.

Furthermore, we conducte both automatic human evaluation and manual human evaluations, comparing ChatGPT to mainstream SOTA models (GECToR and T5) and a widely-used commercial GEC system, Grammarly. These assessments provide valuable and reliable insights into the strengths and weaknesses of ChatGPT for the GEC task. We also conducte an analysis of error types in the document-level GEC, revealing the reasons behind ChatGPT's suboptimal performance in this context.

Based on the results of our experiments and analyses, we summarize the following obervations:

- ChatGPT exhibits a significant disparity with the current SOTA systems in terms of Precision and $F_{0.5}$, while its Recall performance is remarkably superior. This suggests that ChatGPT has the ability to detect errors in the text, but its elevated degree of modification freedom may result in superfluous changes.

- ChatGPT shows significant potential for GEC through zero-shot CoT and few-shot CoT strategies. Even on the JFLEG test set, it has shown minimal difference from the SOTA models and surpassed human evaluations, exhibiting human-like fluency.

- ChatGPT also demonstrates advantages in non-English languages and low-resource environments under the zero-shot CoT strategy, outperforming Transformer models trained from scratch. This highlights the potential of ChatGPT for multilingual GEC tasks.

- Our human evaluations indicate that ChatGPT performs with greater fluency in correcting grammatical errors, albeit with a noticeable

Lan.	Data	#Sents	#Docs	#Doc.len.
EN	CoNLL14	1,312	50	26
	JFLEG test	747	-	-
	BEA19 test	4,477	-	-
	BEA19 dev	-	350	13
	FCE test	-	194	14
DE	Falko-MERLIN	2,337	-	-
ZH	NLPCC18 test	2,000	-	-

Table 1: The datasets used in the evaluation.

tendency towards over-correction. Additionally, as sentence length increases, ChatGPT exhibits a tendency to follow minimal edits and the under-correction rate gradually increases.

- Upon analyzing the error types for document-level GEC, ChatGPT shows relatively poor performance in correcting agreement, coreference, and tense errors across sentences, as well as cross-sentence boundary errors. This could be due to ChatGPT's inherent limitations in processing excessively long sentences.

2 Experimental Setup

2.1 Dataset

Our evaluation involves conducting sentence-level grammatical error correction (GEC) evaluations on a total of five official test sets across three languages: English, German, and Chinese. For English, we select the wildly-used CoNLL14 (Ng et al., 2014) and BEA19 (Bryant et al., 2019) GEC test sets. We use the official test set of NLPCC18 (Zhao et al., 2018) and the official Falko-MERLIN (Boyd et al., 2014) test set for Chinese and German, respectively. The four aforementioned test sets only contain minimal edits that rectify the grammatical errors in a sentence, without necessarily improving fluency or naturalness of the sentence. To assess the error correction abilities of GEC systems more accurately, we also evaluate on the JFLEG (Napoles et al., 2017) test set in English, which represents a range of language proficiency levels and utilizes comprehensive fluency edits to enhance the accuracy of the evaluation. Additionally, we contemplate conducting evaluations on three document-level test sets for the English language, following the methodology suggested by Yuan and Bryant (2021). These test sets include the FCE document-level test set (Yannakoudakis et al., 2011), the BEA19 document-level develop-

ment set ((Bryant et al., 2019)), and the CoNLL14 document-level test set (Ng et al., 2014). Table 1 presents the statistics of the datasets we used.

2.2 Grammatical Error Correction Systems

Currently, among the publicly available GEC systems for sentence-level GEC tasks, the state-of-the-art (SOTA) Seq2Seq GEC model is (m)T5 and its variant models (Rothe et al., 2021), and the SOTA Seq2Edit GEC model is the GECToR model (Omelianchuk et al., 2020). In addition to these SOTA models, we include the Transformer-base (Vaswani et al., 2017) GEC model as our baseline model for comparing the performance of the ChatGPT system. Since TagGEC (Stahlberg and Kumar, 2021) is the SOTA model on the JFLEG test set, we also include it in the comparison. Regarding the comparison of the document-level GEC system, we utilize the MultiEnc-dec model proposed by Yuan and Bryant (2021), which is presently considered as the SOTA for English document-level GEC task.

2.3 ChatGPT system

OpenAI has recently developed several GPT-3.5 series models[2], among which ChatGPT (gpt-3.5-turbo) stands out as the most advanced and specifically optimized for chat functionality. We assess the ChatGPT's performance in the GEC task using official API[3].

2.4 Evaluation Method

Sentence-Level Evaluation GEC systems are evaluated using automated metrics, which compare their output against gold-standard corrections from reference corpora. The selection of automatic evaluation metrics depends on their correlation with human judgments on different types of test sets. Therefore, following previous work (Katsumata and Komachi, 2020; Omelianchuk et al., 2020; Rothe et al., 2021), we utilize the M2 Scorer[4] (Dahlmeier and Ng, 2012) to evaluate the performance of systems on CoNLL14 English, Falko-MERLIN German, and NLPCC18 Chinese GEC tasks. For assessing the BEA2019 test set, we employ an official ERRANT Scorer[5] (Bryant et al., 2019). Additionally, we adopt the GLUE metric[6]

(Napoles et al., 2016) to evaluate the JFLEG test set.

Document-Level Evaluation Based on the approach by Yuan and Bryant (2021), we manually process the raw data to produce document-level references since the available references are only at the sentence level. To evaluate the performance of the ChatGPT system on document-level GEC task, we adopt their method and utilize the official scorer of the BEA19 shared task (Bryant et al., 2019), the ERRANT Scorer (Bryant et al., 2017), for document-level GEC evaluation. For further details, it can be referred to their repository[7].

Human Evaluation Automatic evaluation may introduce bias in reflecting the true performance of a system, especially when there is a limited number of available references. Therefore, we conducte two types of human evaluations, namely automatic human evaluation and manual human evaluation, to more accurately assess the performance of the systems. To conduct automatic human evaluation, we follow the method of Bryant and Ng (2015) and Napoles et al. (2017) to compare the performance of a GEC system against human performance. They measure human performance on the CoNLL14 and JFLEG test sets, respectively, using 10 and 4 sets of human annotations. Furthermore, to further analyze the potential of the ChatGPT system, we invite three annotators to manually annotate the outputs of SOTA GEC systems from various aspects on the CoNLL14 test set. The detailed results are provided in Section 4.

3 Experiments

3.1 The Effect of Zero-Shot Prompts for ChatGPT

Zero-shot Prompts Design We explore the zero-shot capabilities of ChatGPT in GEC from two aspects: zero-shot and zero-shot CoT settings. Table 2 presents the different designed prompting methods. Specifically, for zero-shot, we directly ask the ChatGPT system to `identify and correct any grammatical errors in the sentence while keeping the original sentence structure unchanged as much as possible`. However, we observe that using this normal prompts, ChatGPT tends to generate a multitude of explanations and the format of the resulting answers is often disorganized, necessitating

Method	GEC Prompt
Zero-shot	"role": "system", "content": 'You are an English grammatical error correction tool that can identify and correct grammatical errors in a text.' "role": "user", "content": 'Please identify and correct any grammatical errors in the following sentence while keeping the original sentence structure unchanged as much as possible: '
Zero-shot CoT	"role": "system", "content": 'You are an English grammatical error correction tool that can identify and correct grammatical errors in a text.' "role": "user", "content": 'Please identify and correct any grammatical errors in the following sentence indicated by <input> ERROR </input> tag, you need to comprehend the sentence as a whole before identifying and correcting any errors step by step while keeping the original sentence structure unchanged as much as possible. Afterward, output the corrected version directly without any explanations. Remember to format your corrected output results with the tag <output> Your Corrected Version </output>. Please start: <input> ERROR </input>: '

Table 2: Zero-shot and zero-shot CoT settings for Chat-GPT.

Method	CoNLL14			BEA19 test		
	Pre.	Rec.	$F_{0.5}$	Pre.	Rec.	$F_{0.5}$
Zero-shot	48.5	58.9	50.3	30.5	69.0	34.4
+ CoT	**50.2**	**59.0**	**51.7**†	**32.2**	**70.5**	**36.1**

Table 3: Performance of zero-shot for ChatGPT with CoT prompting methods on the CoNLL14 and BEA19 English test sets. **Bold** values indicate the best Pre/Rec/ $F_{0.5}$ scores. Statistically significant improvements over zero-shot prompting method, as indicated by P_value, $^{\dagger}p < 0.01$. Note that the BEA19 test set is a blind test.

manual intervention. To alleviate this issue, we adopt the zero-shot CoT technique (Kojima et al., 2022) and design our own zero-shot CoT prompt for ChatGPT. Different from the zero-shot prompt, we use special tag <input> Input Sentence </input> to indicate the input sentences, and tell ChatGPT should output the results using the format <output> Your Corrected Version </output>. We instruct ChatGPT to approach the task by comprehending the sentence as a whole before identifying and correcting any errors step by step. Afterward, we request ChatGPT to provide us with the corrected sentences directly, without any explanations. We show some examples of using zero-shot and zero-shot CoT settings for ChatGPT in Appendix A.1.

Preliminary Experimental Results We conduct a preliminary experiment in the CoNLL14 and BEA19 English test sets for ChatGPT to evaluate our designed zero-shot prompting methods, with the results reported in Table 3. We can observe that the **zero-shot + CoT** designed method significantly outperforms the normal **zero-shot** method both on the CoNLL14 and BEA19 test sets. Therefore, the subsequent zero-shot and few-shot experiments in all languages will be combined with the **CoT** prompting method.

3.2 Few-Shot Prompts Design for ChatGPT

Several studies have shown that including few labeled examples in the test input can enhance the performance of LLMs through in-context learning (Brown et al., 2020). In this section, we follow Vilar et al. (2022) to randomly select different in-context labeled examples from development data, as they indicate that different few-shot selection strategies are not necessarily superior to random selection. In particular, we randomly select input-output samples from the CoNLL13 test, BEA19 development, JFLEG development, Falko-MERLIN development, and NLCC18 train sets for the CoNLL14 test, BEA-19 test, JFLEG test, Falko-MERLIN test, and NLPCC18 test sets, respectively. Figure 1 shows the designed GEC few-shot prompts combined with the CoT technique for ChatGPT. In all cases, we test the performance with 1-shot CoT, 3-shot CoT, and 5-shot CoT for all languages.

3.3 ChatGPT Performance on English GEC Task

We evaluate the English GEC task by conducting experiments on the CoNLL4, BAE19 test, and JFLEG test. Our experiment covers various shot paradigms, including zero-shot CoT, 1-shot CoT, 3-shot CoT, and 5-shot CoT.

Tabel 4 shows the overall performance of Chat-GPT with these shot paradigms. We can obtain some interesting observations from the results. Firstly, although there is a large margin between ChatGPT and existing GEC SOAT systems in terms of $F_{0.5}$ and **Precision** scores on the CoNLL14 and BEA19 test sets, ChatGPT surpasses all the previous SOTA systems in **Recall** scores. This indicates that ChatGPT has a higher propensity for error detection and self-correction.

Moreover, ChatGPT performs extremely well on the JFLEG test set. Under the 3-shot CoT strategy,

{"role": "system", "content": 'You are an English/German/Chinese grammatical error correction tool that can identify and correct grammatical errors in a text. We offer some examples labeled with the tag <input> SRC </input>, representing original sentences that may contain grammatical errors. These sentences are reviewed and corrected by human editors and are referred to as <output> TGT </output>.' }
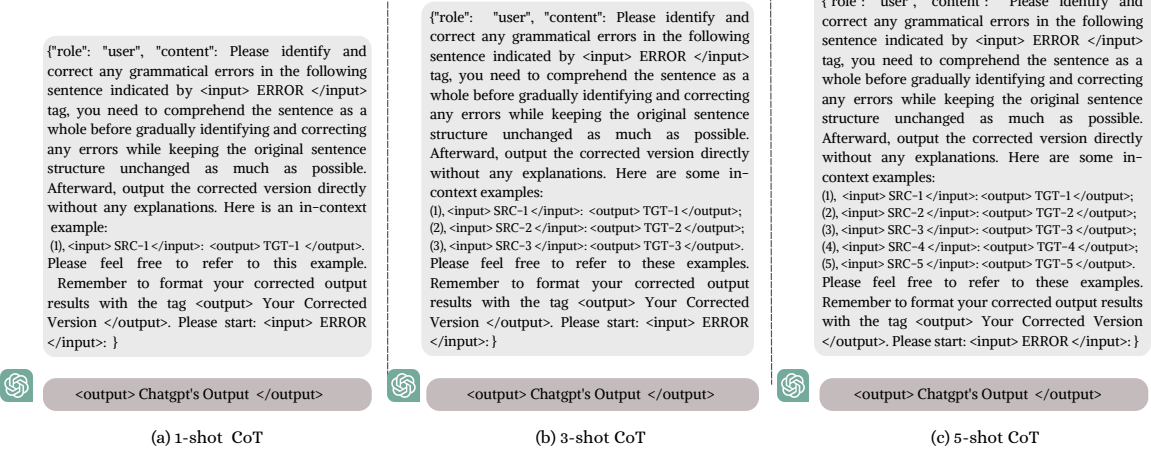
{"role": "user", "content": Please identify and correct any grammatical errors in the following sentence indicated by <input> ERROR </input> tag, you need to comprehend the sentence as a whole before gradually identifying and correcting any errors while keeping the original sentence structure unchanged as much as possible. Afterward, output the corrected version directly without any explanations. Here is an in-context example:
(1), <input> SRC-1 </input>: TGT-1 . Please feel free to refer to this example. Remember to format your corrected output results with the tag Your Corrected Version . Please start: <input> ERROR </input>: }

Chatgpt's Output

(a) 1-shot CoT

{"role": "user", "content": Please identify and correct any grammatical errors in the following sentence indicated by <input> ERROR </input> tag, you need to comprehend the sentence as a whole before gradually identifying and correcting any errors while keeping the original sentence structure unchanged as much as possible. Afterward, output the corrected version directly without any explanations. Here are some in-context examples:
(1), <input> SRC-1 </input>: TGT-1 ;
(2), <input> SRC-2 </input>: TGT-2 ;
(3), <input> SRC-3 </input>: TGT-3 .
Please feel free to refer to these examples. Remember to format your corrected output results with the tag Your Corrected Version . Please start: <input> ERROR </input>: }

Chatgpt's Output

(b) 3-shot CoT

{"role": "user", "content": Please identify and correct any grammatical errors in the following sentence indicated by <input> ERROR </input> tag, you need to comprehend the sentence as a whole before gradually identifying and correcting any errors while keeping the original sentence structure unchanged as much as possible. Afterward, output the corrected version directly without any explanations. Here are some in-context examples:
(1), <input> SRC-1 </input>: <output> TGT-1 </output>;
(2), <input> SRC-2 </input>: TGT-2 ;
(3), <input> SRC-3 </input>: TGT-3 ;
(4), <input> SRC-4 </input>: TGT-4 ;
(5), <input> SRC-5 </input>: TGT-5 .
Please feel free to refer to these examples. Remember to format your corrected output results with the tag Your Corrected Version . Please start: <input> ERROR </input>: }

Chatgpt's Output

(c) 5-shot CoT

Figure 1: An illustration of all few-shot CoT prompts used for ChatGPT to perform Grammatical Error Correction.

System	CoNLL14			BEA19 (test)			JFLEG (test)
	Pre.	Rec.	$F_{0.5}$	Pre.	Rec.	$F_{0.5}$	GLEU
Transformer	60.1	36.6	53.3	60.9	48.3	57.9	55.4
TagGEC (Stahlberg and Kumar, 2021)	72.8	49.5	66.6	72.1	64.4	70.4	**64.7**
GECToR	**75.6**	44.5	66.3	**76.7**	57.8	71.9	58.6
T5 large	72.2	51.4	66.8	73.4	67.0	72.0	62.8
T5 xxl (Rothe et al., 2021)	-	-	**68.9**	-	-	**75.9**	-
ChatGPT (zero-shot CoT)	50.2	59.0	51.7	32.1	**70.5***	36.1	61.4
ChatGPT (1-shot CoT)	52.0	58.1	53.1	34.6	69.7	38.4*	59.7
ChatGPT (3-shot CoT)	51.3	**62.4***	53.2*	34.0	70.2	37.9	63.5*
ChatGPT (5-shot CoT)	50.9	61.8	52.8	32.4	69.9	36.3	62.5

Table 4: Zero-shot CoT and few-shot CoT evaluation results with ChatGPT on the English CoNLL14, BEA19 test, and JFLEG test sets. The comparison GEC models for Transformer, GECToR and T5-large are ones that we trained on the English CLang8 data as Rothe et al. (2021). **Bold** values indicate the best scores across different systems. * denotes the best results among different shot paradigms for ChatGPT.

its **GLEU** score is very close to the SOTA score (falling only 1.2 points short) and it even surpasses the T5 large GEC system by 0.7 points. As we know, JFLEG is a dataset based on fluency editing, which can be used to evaluate a GEC system's ability to correct grammatical errors in a sentence while maintaining fluency. The promising scores obtained by ChatGPT demonstrate its potential for achieving fluency and naturalness in English sentence correction.

Additionally, comparing the performance of different shot CoT strategies on ChatGPT, the few-shot CoT prompting method outperforms the zero-shot CoT prompting method significantly, except for the 1-shot CoT prompting method on the JF-LEG test set. This further demonstrates the effectiveness of the in-context learning method. Interest-

ingly, providing more in-context examples does not necessarily lead to better performance for ChatGPT. Our experiments show that performance tends to decrease when the number of in-context examples exceeds five.

3.4 ChatGPT Performance on non-English and Low-Resource GEC Tasks

We also evaluate non-English GEC tasks on the ChatGPT system by conducting experiments on the German Falko-MERLIN test and Chinese NLPCC18 test sets. It is worth noting that the German training dataset is much smaller than that of English and Chinese, which means that GEC in the German language can be considered a low-resource task. Table 5 presents the evaluation results of zero-shot CoT and few-shot CoT prompt-

System	De (Falko-MERLIN)			Zh (NPLCC18)		
	Pre.	Rec.	$F_{0.5}$	Pre.	Rec.	$F_{0.5}$
Transformer	58.8	34.3	51.5	31.2	20.2	28.1
GECToR	-	-	-	37.4	26.3	34.5
mT5 large	**75.4**	55.1	70.2	**41.5**	25.8	**37.0**
mT5 xxl (Rothe et al., 2021)	-	-	74.8	-	-	-
gT5 xxl (Rothe et al., 2021)	-	-	**76.0**	-	-	-
ChatGPT (zero-shot CoT)	59.9	63.9	60.7	28.7	39.4	28.7*
ChatGPT (1-shot CoT)	61.6	65.6	62.4	25.7	**40.8***	27.8
ChatGPT (3-shot CoT)	63.1	65.3	63.5*	26.4	39.4	28.3
ChatGPT (5-shot CoT)	61.5	**65.9***	62.3	25.2	39.2	27.2

Table 5: Zero-shot CoT and few-shot CoT evaluation results with ChatGPT on the German Falko-MERLIN and Chinese NLPCC18 test sets. The comparison GEC models for Transformer, GECToR and T5-large are ones that we trained on the German CLang8 and Chinese lang8 datasets. **Bold** values indicate the best scores across different systems. * denotes the best results among different shot paradigms for ChatGPT.

ing methods. The results of German and Chinese GEC exhibit similar trends to those in English GEC, where ChatGPT outperforms the SOTA systems in **Recall** scores but scores significantly lower in $F_{0.5}$ and **Precision**.

Furthermore, ChatGPT surpasses the performance of the Transformer base model trained from scratch on the evaluation metrics, suggesting that with proper prompting methods, ChatGPT can effectively perform GEC on low-resource and non-English tasks. This also highlights the potential of ChatGPT for application in multilingual GEC tasks. Another interesting finding is that the performance of zero-shot CoT and few-shot CoT strategies in Chinese is opposite to that in English and German. Specifically, the few-shot CoT prompting method performs comparatively worse than the zero-shot CoT prompting method in Chinese. We hypothesize that, on the one hand, ChatGPT is an LLM centered around English, and German and English belong to the same language family. Therefore, ChatGPT performs similarly on GEC tasks in both languages. On the other hand, the lexicon of Chinese GEC is much more complex than that of English. However, this does not mean that ChatGPT's performance on Chinese GEC is limited to this extent. It may be necessary to design more effective few-shot selection methods to guide ChatGPT and improve its performance.

3.5 Document-Level GEC Performance with ChatGPT

Prior research has shown limited attention toward document-level GEC task, with most studies concentrating on sentence-level GEC tasks using PLMs. As far as we know, only two studies have explored the improvement of performance in English document-level GEC (Chollampatt et al., 2019; Yuan and Bryant, 2021), which only trained on CNN/Transformer architecture from scratch with their approach. In this section, we evaluate English document-level GEC tasks on the ChatGPT system by conducting experiments on the CoNLL14 document-level test, FCE document-level test, and BEA19 document-level development sets followed by Yuan and Bryant (2021). Regarding the design of prompts, we simply replace "sentence" with "document" in Table 2 and Figure 1.

Table 6 presents the evaluation results of the zero-shot CoT prompting method for ChatGPT. Similar to sentence-level GEC tasks, ChatGPT exhibits good performance in terms of **Recall**, but its scores in $F_{0.5}$ and **Precision** are significantly lower in the context of document-level GEC models that are trained from scratch on Transformer without leveraging knowledge from large language models. It is worth noting that this poor performance raises concerns and warrants further investigation.

Furthermore, the large number of sentences in each document has impeded our exploration of the few-shot CoT prompting method. To address this issue, we only conduct 1-shot CoT experiments using the BEA19 development set, which has the

System	CoNLL14			FCE (test)			BEA19 (dev)		
	Pre.	Rec.	$F_{0.5}$	Pre.	Rec.	$F_{0.5}$	Pre.	Rec.	$F_{0.5}$
SingEnc	59.8	27.3	48.3	61.6	45.0	57.4	57.0	43.2	53.5
MultiEnc-enc	63.2	28.0	50.5	65.6	42.7	59.2	62.1	41.7	56.5
MultiEnc-dec	**64.6**	28.7	**51.6**	**65.4**	44.2	**59.7**	**62.6**	40.7	**56.6**
ChatGPT(zero-shot CoT)	42.3	**40.8**	42.0	46.3	**49.9**	47.0	47.7	**50.8**	48.3

Table 6: Zero-shot CoT evaluation results with ChatGPT on the English CoNLL14 document-level test, FCE document-level test, and BEA19 document-level development sets. The comparison document-level GEC results for SingEnc, MultiEnc-enc, and MultiEnc-dec models are reported by Yuan and Bryant (2021), where MultiEnc-dec is the current SOTA model. **Bold** values indicate the best scores across different systems.

Method	BEA19 (dev)		
	Pre.	Rec.	$F_{0.5}$
ChatGPT (zero-shot CoT)	**47.7**	50.8	**48.3**[†]
ChatGPT (1-shot CoT)	42.5	**55.6**	44.6

Table 7: Performance of zero-shot CoT and 1-shot CoT prompting methods for ChatGPT on the BEA19 document-level development set. **Bold** values indicate the best scores. Statistically significant improvement over the 1-shot CoT prompting method is reported using P_value, [†]$p < 0.01$.

System	CoNLL14	JFLEG
	$F_{0.5}$	GLEU
Human	72.58	62.37
Transformer	66.97	55.41
GECToR	80.49	58.61
T5 large	**81.19**	62.82
ChatGPT (0-shot)	69.74	61.42
ChatGPT (1-shot)	71.55	59.65
ChatGPT (3-shot)	71.73*	**63.52***
ChatGPT (5-shot)	70.66	62.53

Table 8: Zero-shot CoT and few-shot CoT Prompting methods for ChatGPT in comparison to automatic human evaluation performance. **Bold** values indicate the best scores across different systems. * denotes the best results among different shot paradigms for ChatGPT.

fewest average number of sentences compared to the other two test sets. For in-context examples, we randomly select one sample from the FCE test set. As shown in Table 7, it appears that the 1-shot CoT strategy is not effective for document-level GEC. While there is a noticeable increase in **Recall**, the $F_{0.5}$ and **Precision** scores still decreased significantly. In light of these results, we speculate that ChatGPT may not be sufficiently capable of processing long sentences, which often require high levels of coherence and consistency between sentences. In the subsequent analysis section 4.3, we will examine ChatGPT's shortcomings from the perspective of correcting different types of errors.

4 Human Evaluation and Analysis

4.1 Automatic Human Evaluation

Bryant and Ng (2015) is the first to attempt to measure human performance on the CoNLL14 test set using 10 references. Specifically, they calculate the performance of each annotator by comparing its corrections to the other 9 annotators. The average of the 9 $F_{0.5}$ scores was taken as the final score for the human-level performance. To perform an automatic human evaluation on the ChatGPT performance, we adopt the approach used by Bryant

and Ng (2015) and Napoles et al. (2017) to compare the performance of different GEC systems with that of humans on the CoNLL14 and JFLEG test sets. The reported $F_{0.5}$ scores for human performance on these test sets are 72.58 and 62.37, respectively.

Table 8 shows the automatic human evaluation results. We have observed some interesting phenomena. Firstly, the $F_{0.5}$ score of ChatGPT on the CoNLL14 test is lower than that of the Transformer base model in Table 4, but the human evaluation shows a significant improvement for ChatGPT. This suggests that the existing evaluation methods may underestimate the performance of ChatGPT on sentence-level GEC tasks. Although the best few-shot CoT strategy for ChatGPT achieved a human evaluation score far lower than the scores of two mainstream SOTA models, it is only 0.85 $F_{0.5}$ points lower than the human-level evaluation score, indicating that ChatGPT has great potential in the
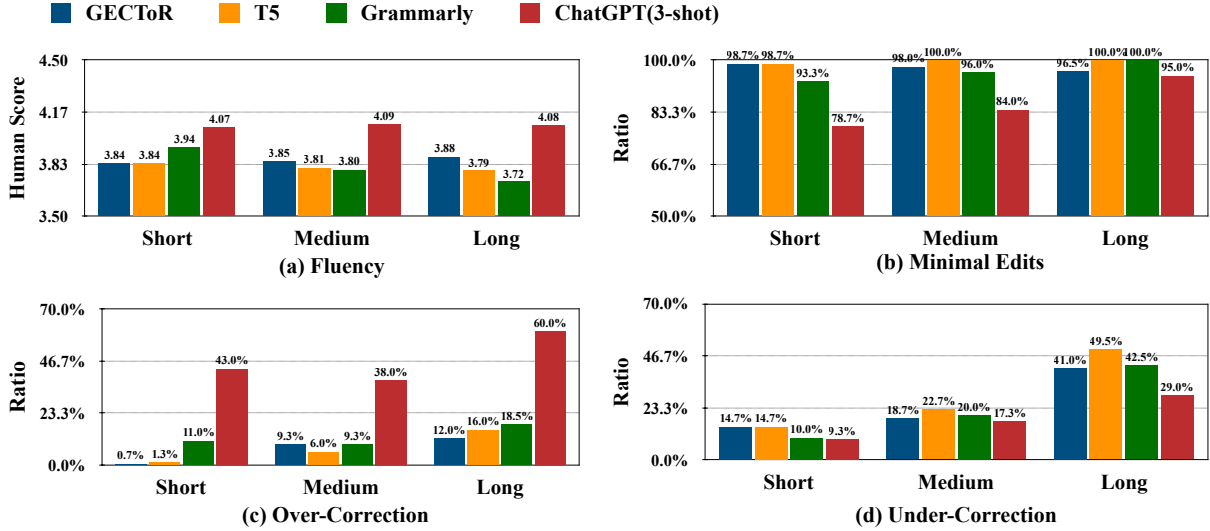
Figure 2: The results of manual human evaluations on four criteria: Fluency, Minimal Edits, Over-Correction, and Under-Correction. The statistical analysis is based on the average scores provided by three evaluators.

sentence-level GEC task.

Furthermore, the human evaluation performance of ChatGPT on the JFLEG test set is impressive. The best few-shot CoT prompting method not only exceeds the human-level evaluation score by 1.15 GLEU points but also outperforms the strong baseline T5 large model by 0.7 GLEU points. This observation suggests that the sentences corrected by ChatGPT exhibit a high level of fluency and naturalness.

4.2 Manual Human Evaluation

To explore ChatGPT's true performance on the CoNLL14 test set, we carry out a manual human evaluation. In addition to comparing with two mainstream SOTA GEC systems (GECToR, T5 large), we also consider the widely-used commercial system, Grammarly[8], renowned for its ability to detect and correct a range of errors in English texts, such as spelling, punctuation, grammar, and word choice, using features like Correctness, Clarity, Engagement, and Delivery. In this evaluation, we only use its free and open Correctness feature to correct sentences. We select the 50 longest, 50 medium-length, and 50 shortest sentences from the CoNLL14 test set and invite three postgraduate students with international study experience to conduct a manual evaluation. Our reference is the widely-used *official-2014.combined.m2* version (Ng et al., 2014) of CoNLL14, and we obtain the final evaluation by averaging the scores from

---

the three students. The evaluation focus on four main categories: Fluency, Minimal Edits, Over-Correction, and Under-Correction. Some examples are shown in Appendix A.2.

Fluency refers to ensuring that the corrected sentence conforms to the linguistic conventions, is easily comprehensible, flows naturally, and preserves the intended meaning of the original sentence while correcting the grammatical errors. We define a 1-5 rating scale, where 1 represents the lowest and 5 is the highest level of fluency. From the results shown in Figure 2(a), it can be seen that ChatGPT exhibits significantly better fluency in correcting grammar errors of short, medium, and long sentences compared to the other three mainstream systems. In addition, we also observed that among the three mainstream systems, Grammarly performs more fluently in correcting grammar errors in short sentences, while GECToR performs better in long sentences. As for sentences of medium length, the three systems perform similarly.

Minimal Edits The CoNLL14 test set includes grammatically incorrect sentences that have been created by making minimal edits to grammatically correct sentences. We compare the output of ChatGPT and three other systems to the references sentences in order to determine whether they use minimal edits. The annotators assign a value of 1 to indicate that a sentence followed minimal edits, and 0 to indicate that it does not. We compute the proportion of minimal edits for sentences of different lengths. The results are shown in Fig-

ure 2(b). Compared to the other three systems, ChatGPT is not particularly inclined to use minimal edits, especially in short sentences. However, as the length of the sentence increases, ChatGPT becomes more willing to follow minimal edits. Furthermore, Grammarly exhibits a lower level of conformity to minimal edits in short sentences compared to two mainstream SOTA systems. In terms of medium and long sentences, GECToR shows a lower level of minimal edits compared to T5 large. Interestingly, comparing Figure 2(a), we find that the less a system adheres to minimal edits, the better the fluency performance of the generated sentences.

Over-Correction Since there can be multiple ways to correct a sentence, sometimes a correction that differs from the references does not necessarily mean the sentence has not been corrected. Over-Correction is used to evaluate a system's ability to produce correct results beyond what is indicated by the reference correction. We also ask the annotators to assign a value of 1 to indicate that a sentence is an over-correction, and 0 to indicate that it is not. We calculate the proportion of Over-Correction for sentences with varying lengths and the results are depicted in Figure 2(c). ChatGPT surpasses the other three GEC systems in its capacity to over-correct sentences of different lengths, which indicates that it can correct sentences freely and diversely, which is consistent with its good fluency demonstrated in Figure 2(a). Additionally, we observe that Grammarly also exhibits over-correction compared to GECToR and T5 large models.

Under-Correction refers to cases where a GEC system fails to correct certain grammatical errors that exist in the original sentence. Our requirement for annotators is to check any errors present in the generated sentences by GEC systems that are not identified or corrected. Similar to Over-Correction, 1 indicates a sentence is an under-correction, and 0 indicates that it is not. Interestingly, in contrast to Over-Correction, ChatGPT demonstrates fewer Under-Corrections than the other systems in short, medium, and long sentences, as shown in Figure 2(d). This suggests that ChatGPT has great potential for GEC tasks. Furthermore, T5 large GEC system is prone to Under-Corrections, which highlights that a higher $F_{0.5}$ score does not imply a more perfect GEC system. We also observe that as sentence length increases, ChatGPT exhibits a stronger tendency towards Under-Correction. This

may explain its poor performance in document-level GEC tasks.

Through manual analysis, we can conclude that ChatGPT generates grammatically corrected sentences that are highly fluent, which may be attributed to its powerful diversity and free-generation ability.

4.3 Fine-grained Error Analysis for Document-Level GEC

To evaluate ChatGPT's ability to correct various types of errors in document-level GEC, we conduct an study based on Yuan and Bryant (2021), using the ERRANT toolkit (Bryant et al., 2017) to analyze the results on the BEA19 document-level development set. The results are presented in Table 9, comparing the performance of ChatGPT and the MultiEnc-dec document-level GEC system on POS-based fine-grained error types. ChatGPT demonstrates strong performance in addressing errors related to punctuation, nouns, and possessive nouns, achieving improvements of **6.0**, **7.1**, and **0.8** $F_{0.5}$ scores, respectively. Regarding errors that frequently demand a strong understanding of agreement, coreference, or tense across multiple sentences, these are typically considered document-level errors. The ChatGPT performs significantly worse than the MultiEnc-dec model in correcting errors related to subject-verb agreement, prepositions, noun numbers, determiners, and pronouns (e.g. VERB:SVA **-5.1** $F_{0.5}$, VERB:TENSE **-3.7** $F_{0.5}$, VERB:FORM **-15.5** $F_{0.5}$, NOUN:NUM **-6.0** $F_{0.5}$, PRON **-7.7** $F_{0.5}$.). Moreover, ChatGPT's performance was even worse in handling cross-sentence boundary errors, as evidenced by a decline of **14.8** $F_{0.5}$ in CONJ and **18.3** $F_{0.5}$ in PUNCT. We speculate the underperformance of the ChatGPT could be attributed to its potential limitations in terms of contextual memory, consistency, and coherence across sentences at the document-level.

5 Conclusion

In this study, we undertake a thorough examination of ChatGPT's performance and potential in GEC, as well as compare it with current SOTA models. To the best of our knowledge, we are the first to design zero-shot CoT and few-shot CoT settings for ChatGPT in the GEC task. Our experiments involve evaluating ChatGPT on five official test sets in three different languages, as well as on three document-level GEC test sets in En-

Error-Type	MultiEnc-dec			ChatGPT(zero-shot)			ChatGPT(1-shot)			Differ.
	Pre.	Rec.	$F_{0.5}$	Pre.	Rec.	$F_{0.5}$	Pre.	Rec.	$F_{0.5}$	$F_{0.5}$
ADJ	44.4	15.6	**32.5**	34.9	20.2	30.4	32.2	32.5	32.3	-0.2
ADJ:FORM	100.0	25.0	62.5	60.0	75.0	62.5	56.5	81.3	60.2	0.0
ADV	42.1	17.9	**33.2**	18.7	17.7	18.5	18.6	27.7	19.9	-13.3
CONJ	46.2	14.3	**31.9**	13.2	22.5	14.4	15.0	37.5	17.1	-14.8
CONTR	85.0	58.6	**78.0**	38.7	38.7	38.7	28.6	51.6	31.4	-39.3
DET	63.7	51.3	**60.8**	57.9	59.1	58.1	54.1	66.0	56.1	-2.7
MORPH	70.6	33.3	**57.7**	44.6	55.6	46.5	45.2	65.6	48.2	-9.5
NOUN	38.1	13.3	27.7	41.2	21.5	**34.8**	31.5	26.9	30.4	**+7.1**
NOUN:INFL	100.0	75.0	**93.8**	88.9	72.7	85.1	90.9	90.9	90.9	-2.9
NOUN:NUM	74.2	47.6	**66.7**	59.1	60.2	59.3	58.1	74.5	60.7	-6.0
NOUN:POSS	63.0	51.8	60.4	68.4	65.0	67.7	69.0	66.7	**68.5**	**+0.8**
ORTH	71.2	59.9	**68.6**	64.3	76.1	66.4	63.3	78.6	65.8	-2.2
OTHER	38.1	22.5	**33.5**	26.5	29.8	27.1	18.8	33.5	20.6	-6.4
PART	58.5	40.7	53.8	71.4	39.7	**61.6**	60.3	55.6	59.3	-7.8
PREP	64.7	41.9	**58.3**	56.6	43.2	53.3	49.4	52.1	49.9	-5.0
PRON	55.7	43.1	**52.6**	43.1	53.9	44.9	33.6	56.9	36.6	-7.7
PUNCT	70.1	47.3	**63.9**	42.2	63.2	45.2	42.8	62.2	45.6	-18.3
SPELL	86.2	53.9	76.9	81.5	89.3	**82.9**	81.3	88.6	82.7	**+6.0**
VERB	44.8	21.6	**36.9**	39.5	21.9	34.0	31.5	31.2	31.5	-2.9
VERB:FORM	71.1	60.9	**68.8**	50.5	68.6	53.3	45.5	68.6	48.8	-15.5
VERB:INFL	80.0	66.7	76.9	80.0	66.7	76.9	57.1	66.7	58.8	0.0
VERB:SVA	72.4	74.5	**72.8**	64.7	82.9	67.7	61.8	84.3	65.3	-5.1
VERB:TENSE	63.5	44.8	**58.6**	58.1	44.5	54.8	55.4	53.2	54.9	-3.7
WO	64.7	37.5	**56.5**	41.0	46.6	42.0	36.2	47.7	38.0	-14.5
Total	62.6	40.7	**56.6**	47.7	50.8	48.3	42.5	55.6	44.63	-8.3

Table 9: Fine-grained error types performance of **ChatGPT** and **MultiEnc-dec** systems on the BEA19 development set. **Differ.** refers to the difference in $F_{0.5}$ scores between the two systems. **Bold** values indicate the best $F_{0.5}$ scores.

glish. The experimental results demonstrate that ChatGPT has strong error detection capabilities and can generate sentences with human-like fluency, despite its poor performance in precision and $F_{0.5}$ scores. Additionally, we find that ChatGPT also demonstrate its advantages in multilingual and low-resource settings, showing great potential for multilingual GEC. The results of further human evaluations once again confirm that ChatGPT is a highly fluent GEC system, and demonstrate that its performance can be better leveraged through the use of chain-of-thought prompts. However, our analysis of ChatGPT's ability to correct various types of errors in document-level GEC indicates that it performs poorly on most error types, such as agreement, coreference, tense errors across sentences, and cross-sentence boundary errors.

Limitations

Our study has some limitations as follows:

- Our current work solely focus on exploring ChatGPT's potential performance in GEC, without delving into investigating other mainstream LLMs, such as other GPT-3.5 series models or the latest GPT-4 model from OpenAI. However, we plan to include these models in our future work.

- Our study highlights that ChatGPT exhibits higher fluency and tends to make more casual edits in GEC tasks. However, in certain practical applications, such as language education, users may require only minimal editing corrections (Coyne and Sakaguchi, 2023). Therefore, exploring ChatGPT's potential in this aspect through the design of prompting methods is also a worthwhile pursuit.

- For the few-shot settings using in-context learning, we follow Vilar et al. (2022) in randomly selecting examples from the development set,

without specifically designing corresponding examples for each sentence. It is worth investigating whether selecting better examples could help improve ChatGPT's performance in GEC.

Acknowledgments

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

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. Curran Associates, Inc.

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, Florence, Italy. Association for Computational Linguistics.

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

Christopher Bryant 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 (Volume 1: Long Papers)*, pages 697–707, Beijing, China. Association for Computational Linguistics.

Shamil Chollampatt, Weiqi Wang, and Hwee Tou Ng. 2019. Cross-sentence grammatical error correction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 435–445, Florence, Italy. Association for Computational Linguistics.

Steven Coyne and Keisuke Sakaguchi. 2023. An analysis of gpt-3's performance in grammatical error correction.

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

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Satoru Katsumata and Mamoru Komachi. 2020. Stronger baselines for grammatical error correction using a pretrained encoder-decoder model. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 827–832, Suzhou, China. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Aiwei Liu, Xuming Hu, Lijie Wen, and Philip S. Yu. 2023. A comprehensive evaluation of chatgpt's zero-shot text-to-sql capability.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2016. GLEU without tuning. *eprint arXiv:1605.02592 [cs.CL]*.

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: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

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, Baltimore, Maryland. Association for Computational Linguistics.

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, Seattle, WA, USA → Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *arxiv preprint*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver?

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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 (Volume 2: Short Papers), pages 702–707, Online. Association for Computational Linguistics.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization.

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, Portland, Oregon, USA. Association for Computational Linguistics.

Zheng Yuan and Christopher Bryant. 2021. Document-level grammatical error correction. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 75–84, Online. Association for Computational Linguistics.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018. Overview of the nlpcc 2018 shared task: Grammatical error correction. In *Natural Language Processing and Chinese Computing*, pages 439–445, Cham. Springer International Publishing.

A Appendix

A.1 Examples of zero-shot settings

Figure 3 presents an example of the zero-shot setting for ChatGPT. Figure 4 and Figure 5 show the examples of zero-shot CoT settings for ChatGPT.

A.2 GEC System Output Examples

Table 10 presents some output examples generated by various system. It is noticeable that ChatGPT, in short and medium-length sentences, does not adhere strictly to the minimum editing principle and tends to over-correct. However, the resulting sentences exhibit higher fluency. In the case of long sentences, ChatGPT also demonstrates instances of over-correction, such as replacing "while" with "and." While there is no grammatical error, the absence of a conspicuous contrasting relationship results in a decline in the overall sentence quality.

Example of zero-shot setting

YJ Please identify and correct any grammatical errors in the following sentence while keeping the original sentence structure unchanged as much as possible: Therefore I think a know genetic risk should not be obligated to tell his or her family or relatives.

Therefore, I think that someone who knows they have a genetic risk should not be obligated to tell their family or relatives.

The corrected sentence maintains the original structure while correcting the following errors:

- "a know" should be "someone who knows"
- "his or her" should be "their" for gender neutrality
- "obligated" should be followed by "to" in order to form a complete infinitive verb phrase.

More explanation

Figure 3: Zero-shot setting for ChatGPT.

Example of zero-shot CoT setting 1

YJ Please identify and correct any grammatical errors in the following sentence indicated by <input> ERROR </input> tag, you need to comprehend the sentence as a whole before identifying and correcting any errors step by step while keeping the original sentence structure unchanged as much as possible. Remember to format your corrected output results with the tag <output> Your Corrected Version </output>. Please start: <input> Therefore I think a know genetic risk should not be obligated to tell his or her family or relatives. </input>:

First, "know" should be changed to "known" to match the past tense "should not be obligated."

Second, the sentence contains a gendered pronoun ("his or her") and then switches to a non-gendered pronoun ("their"). To maintain consistency, we can replace "his or her" with "their."

With those changes, the corrected sentence is:

Chain-of-thought reasoning process

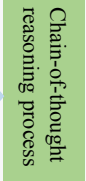Therefore, I think someone with a known genetic risk should not be obligated to tell their family or relatives.

Figure 4: Zero-shot CoT setting for ChatGPT. The prompts without uing "afterward, output the corrected version directly without any explanations".

Example of zero-shot CoT setting 2

YJ — Please identify and correct any grammatical errors in the following sentence indicated by <input> ERROR </input> tag, you need to comprehend the sentence as a whole before identifying and correcting any errors step by step while keeping the original sentence structure unchanged as much as possible. Afterward, output the corrected version directly without any explanations. Remember to format your corrected output results with the tag Your Corrected Version . Please start: <input> Therefore I think a know genetic risk should not be obligated to tell his or her family or relatives. </input>:

Therefore, I think someone with a known genetic risk should not be obligated to tell their family or relatives.

ChatGPT's output

Figure 5: Zero-shot CoT setting for ChatGPT.

Type	System	Examples	Human
S	Src	It can make people become closer as well.	
	Ref.	It can make people become closer as well.	
	GECToR	It can make people become closer as well.	4 | 1 | 0 | 0
	T5 large	It can make people become closer as well.	4 | 1 | 0 | 0
	Grammarly	It can make people become closer as well.	4 | 1 | 0 | 0
	ChatGPT(3-shot)	It can bring people closer together as well.	5 | 0 | 1 | 0
M	Src	Recent research has shown that the more people spend time on social media sites, the less they become ambitious.	
	Ref.	Recent research has shown that the more people spend time on social media sites, the less they become ambitious.	
	GECToR	Recent research has shown that the more people spend time on social media sites, the less they become ambitious.	4 | 1 | 0 | 0
	T5 large	Recent research has shown that the more people spend time on social media sites, the less they become ambitious.	4 | 1 | 0 | 0
	Grammarly	Recent research has shown that the more people spend time on social media sites, the less they become ambitious.	4 | 1 | 0 | 0
	ChatGPT(3-shot)	Recent research has shown that the more time people spend on social media sites, the less ambitious they become.	5 | 0 | 1 | 0
L	Src	Though it is said that genetic testing involves emotional and social risks due to the test results, while the potential negative impacts of the risk still exist, the consequence will be significant if other members of his or her family do not know.	
	Ref.	Though it is said that genetic testing involves emotional and social risks due to the test results, while the potential negative impacts of the risk still exist, the consequences will be significant if other members of his or her family do not know.	
	GECToR	Though it is said that genetic testing involves emotional and social risks due to the test results, while the potential negative impacts of the risk still exist, the consequences will be significant if other members of his or her family do not know.	4 | 1 | 0 | 0
	T5 large	Though it is said that genetic testing involves emotional and social risks due to the test results, while the potential negative impacts of the risk still exist, the consequence will be significant if other members of his or her family do not know.	4 | 1 | 0 | 0
	Grammarly	Though it is said that genetic testing involves emotional and social risks due to the test results, while the potential negative impacts of the risk still exist, the consequence will be significant if other members of his or her family do not know.	4 | 1 | 0 | 1
	ChatGPT(3-shot)	Though it is said that genetic testing involves emotional and social risks due to the test results, and the potential negative impacts of the risk still exist, the consequences will be significant if other members of his or her family do not know.	3 | 1 | 1 | 0

Table 10: Different GEC systems output examples. The words highlighted in red indicate grammatical errors, while those highlighted in blue represent the corrected version by ChatGPT. **Human** denotes the evaluator's scores for Fluency | Minimal Edits | Over-Correction | Under-Correction. **S/ M/ L** stands for short/ medium/ long sentences.