Rethinking the Roles of Large Language Models in Chinese Grammatical Error Correction

Anonymous ACL submission

Abstract

Recently, Large Language Models (LLMs) have been widely studied by researchers for their roles in various downstream NLP tasks. As a fundamental task in the NLP field, Chinese Grammatical Error Correction (CGEC) aims to correct all potential grammatical errors in the input sentences. Previous studies have shown that LLMs' performance as correctors on CGEC remains unsatisfactory due to its challenging task focus. To promote the CGEC field to better adapt to the era of LLMs, we rethink the roles of LLMs in the CGEC task so that they can be better utilized and explored in CGEC. Considering the rich grammatical knowledge stored in LLMs and their powerful semantic understanding capabilities, we utilize LLMs as explainers to provide explanation information for the CGEC small models during error correction to enhance performance. We also use LLMs as evaluators to bring more reasonable CGEC evaluations, thus alleviating the troubles caused by the subjectivity of the CGEC task. In particular, our work is also an active exploration of how LLMs and small models better collaborate in downstream tasks. Extensive experiments [1] and detailed analyses on widely used datasets verify the effectiveness of our intuition and the proposed methods.

1 Introduction

Large Language Models (LLMs) are undoubtedly the hottest stars in the AI and NLP community. Due to the unified paradigm for various tasks and amazing emergent ability, more and more researchers and works have begun to focus on how to better apply LLMs to downstream task scenarios, such as sequence understanding (Yu et al., 2023), financial analysis (Wu et al., 2023), and medical healthcare (Wang et al., 2023).

In the vast field of Chinese NLP research, Chinese Grammatical Error Correction (CGEC) has long been regarded as a fundamental task (Ma et al., 2022). The CGEC task aims to correct all possible grammatical errors in the input sentence, which is challenging because it requires the models to have a comprehensive understanding ability for the complex semantics of the text. In the era of LLMs, some works have explored the possibility of LLMs for CGEC (Fang et al., 2023; Li et al., 2023b). Their consensus is that even with supervised fine-tuning on CGEC data, the performance of LLMs on the CGEC task is still unsatisfactory. The main reason is that the relatively free generation paradigm makes the sentences generated by LLMs often unable to meet the minimum change principle pursued by CGEC. Therefore, adapting and applying LLMs in the CGEC field have encountered a stagnant dilemma.

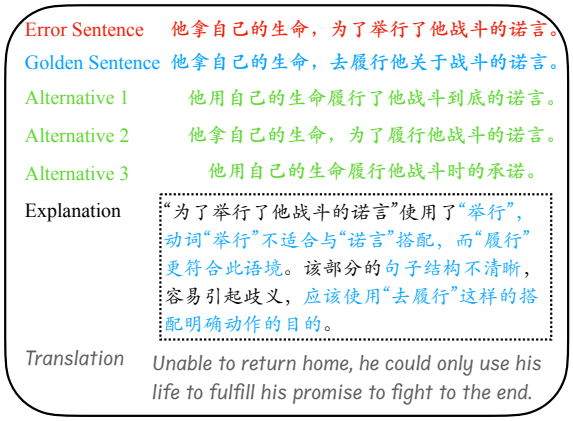

Figure 1: The example of subjectivity and explainability of CGEC. The explanation is produced by ChatGPT.

To address this dilemma, our work rethinks the proper utilization of LLMs to promote the development of the CGEC field. Overviewing recent GEC research trends, the subjectivity and explainability of GEC have received great attention (Ye et al., 2023; Song et al., 2023; Kaneko and Okazaki, 2023a). As illustrated in Figure 1, a grammatically incorrect sentence often has different cor-

[1] Our code will be made public after peer review.

rection methods to keep its meaning unchanged and its grammar correct. Therefore, enabling evaluators to perform comprehensively and flexibly has always been an unsolved challenge. In addition, we also see from Figure 1 that the explanation of the wrong sentence contains instructive information and knowledge for error correction. If we can obtain high-quality explanations of wrong sentences, it can undoubtedly improve the CGEC performance. The basis for high-quality explanations of ungrammatical sentences is rich grammatical knowledge, while flexible CGEC evaluation requires the evaluator to have comprehensive semantic understanding capabilities. Intuitively, for LLMs, the massive training corpus gives them **sufficient grammatical knowledge**, and the emergence phenomenon gives them **excellent semantic understanding capabilities.** More importantly, the two processes of explanation and evaluation are not restricted by the minimum change principle, and they can give enough free space to the generation paradigm of LLMs.

Motivated by the above intuitions, we believe that LLMs can be leveraged to provide high-quality explanations and accurate evaluations for small CGEC models. Therefore, we propose an **EX**planation-**A**ug**M**ented training framework (**EXAM**) and a **SE**mantic-incorporated **E**valuation framework (**SEE**) for CGEC based on LLMs. Specifically, (1) EXAM mines broad explanation information (including error types, reference corrections, and error explanations) related to grammatically incorrect sentences from LLMs, and then utilizes mined information to enhance the training of small models, thereby ultimately improving the CGEC performance of small models. (2) SEE requires LLMs to balance the edits annotated in the golden data with the evaluated model's edits that do not alter the original semantics of the input sentence. This ensures more accurate and comprehensive evaluation results that consider both grammar and semantics.

Extensive experiments and detailed analyses demonstrate the effectiveness and competitiveness of our proposed methods. In summary, our technical contributions and impacts are in three folds:

- We propose EXAM, which utilizes LLMs as the explainer to enhance the training of small models, and SEE, which aims to empower the evaluation of more subjective CGEC tasks through the intervention of LLMs.
- **For CGEC field**, we reposition the roles of LLMs to give full play to the strengths of LLMs and promote the adaptation of LLMs to the CGEC task.
- **For LLMs community**, our work explores collaborative cooperation between LLMs and small models on downstream tasks and, to a certain extent, reveals how LLMs and small models coexist and prosper in the future.

2 Related Work

In the era of LLMs, considering the superior performance of LLMs (Liu et al., 2023; Li et al., 2023a), researchers have invested lots of energy in studying LLMs for GEC tasks.

First, some works evaluate LLMs on GEC (Fang et al., 2023; Penteado and Perez, 2023; Qu and Wu, 2023; Li et al., 2023b; Kwon et al., 2023; Davis et al., 2024). In general, GEC-related tasks are challenging for LLMs. There are many reasons for this challenge, such as the inconvenience caused to LLMs by the minimum change principle. To address the challenges, some researchers also focus on training LLMs on GEC data (Fan et al., 2023; Zhang et al., 2023; Su et al., 2023). Still unsatisfactory, even after supervised fine-tuning, the performance of LLMs still cannot prove that LLMs have fully adapted to the GEC field. For example, the $F_{0.5}$ scores reported by GrammarGPT (Fan et al., 2023) still do not exceed 40.0. As a result, researchers begin to pay attention to whether LLMs can have other roles in the GEC field, instead of directly acting as the corrector. Kaneko and Okazaki (2023b) propose to improve the GEC performance by letting LLMs predict edit spans. Östling et al. (2023) and Sottana et al. (2023) explore the potential of using LLMs as evaluators for English and Swedish GEC tasks. Song et al. (2023) and Kaneko and Okazaki (2023a) propose the new task of grammar error explanation and have proved the ability of LLMs to explain grammatical error. However, they do not go further to utilize the explanation information in training GEC models. *To the best of our knowledge, our work is the first to comprehensively think about and design how to make full use of LLMs in the training and evaluation process of GEC small models. More importantly, our work rethinks how LLMs and small models should coexist and progress together in the era of LLMs, contributing their respective strengths to the advancement of downstream tasks.*

3 Motivation and Methodology

3.1 Motivation

Minimum Change Principle In the long-term GEC or CGEC research, the setting followed by researchers is the "minimum change principle", that is, an ideal model should be able to convert grammatically wrong sentences into correct sentences with minimal changes or editing costs. However, with the development of deep learning and Pre-trained Language Models, the enhancement of model capabilities has conflicted with this principle because it limits the model's space for self-development to a certain extent. Especially with the emergence of LLMs, the performance obtained by directly using LLMs to complete the GEC task is not satisfactory. Many observations and empirical results indicate that the key reason for the unsatisfactory performance of LLMs on CGEC is that the relatively freer text generation mode of LLMs is unsuitable for the GEC task. For example, LLMs often produce sentences that are grammatically correct and semantically consistent with the erroneous input sentence, but the literal text differs significantly from the input sentence. This situation often fails in traditional evaluation metrics, resulting in the low performance of LLMs.

LLMs as Explainer Given the limitations of directly employing LLMs as correctors due to the minimum change principle, can we adopt an alternative approach to leverage LLMs more effectively for CGEC and circumvent the constraints imposed by this principle? First, let's consider what humans do when they encounter grammatical errors, particularly when they are unsure how to correct them. The most direct and effective solution is to turn to a teacher or grammar reference book. Then, the teacher or reference book would give specific explanations or reasons for grammatical errors to help humans make corrections successfully. **Drawing inspiration from human actions, why can't we consider LLMs as explainers similar to teachers or reference books?** As mentioned in the previous paragraph, the fact that LLMs can generate grammatically correct sentences means that LLMs store rich grammatical knowledge. Therefore, we believe that if explanations related to error sentences can be obtained from LLMs and utilized in the training of small models, then these explanations embodying grammatical knowledge from LLMs can definitely enhance the performance of small models. In particular, the role of LLMs as explainers does not need to be limited by the minimum change principle, and it is a simple yet effective process for LLMs to use their own grammatical knowledge to explain wrong sentences.

LLMs as Evaluator Considering the subjective nature of the CGEC task, a sentence with grammatical errors often has different correction methods. We argue that the ideal evaluation that can truly reflect the CGEC performance should consider the correction results given by the model as comprehensively as possible. As long as the model gives a sentence that is consistent with the original semantics of the wrong sentence and has no grammatical errors, then its correction should be considered successful. Suppose we want to achieve this ideal evaluation from the perspective of dataset construction. In that case, we need to manually annotate the dataset with as many correct reference sentences corresponding to the wrong sentences as possible. However, such an annotation process is expensive and time-consuming. Even though there are already multi-reference datasets such as MuCGEC (Zhang et al., 2022), we still believe that automatic evaluation based on such datasets is not flexible enough because the fixed reference correct sentences of the dataset are still limited after all. **Motivated by the process of teachers correcting students' sentences with grammatical errors, why can't we utilize LLMs as evaluators to play the role of a teacher reviewing grammatical errors?** Intuitively, LLMs not only store rich grammatical knowledge but also have an excellent ability to perceive text semantics. Therefore, we believe that they are fully qualified to be flexible and excellent teachers (i.e., evaluators) who review the answers of models in the GEC task.

3.2 Explanation-Augmented Training

As introduced in the above section, we propose the **EX**planation-**A**ug**M**ented training framework (**EXAM**) (as illustrated in Figure 2) to mine explanation information and grammatical knowledge from LLMs and inject them into small models, ultimately achieving the purpose of using LLMs to enhance the performance of small models. Based on our understanding of the CGEC task, we divide the explanation information (note that the "explanation" we consider here is the LLMs analysis of wrong sentences in a broad sense) we want to obtain from LLMs into three categories:

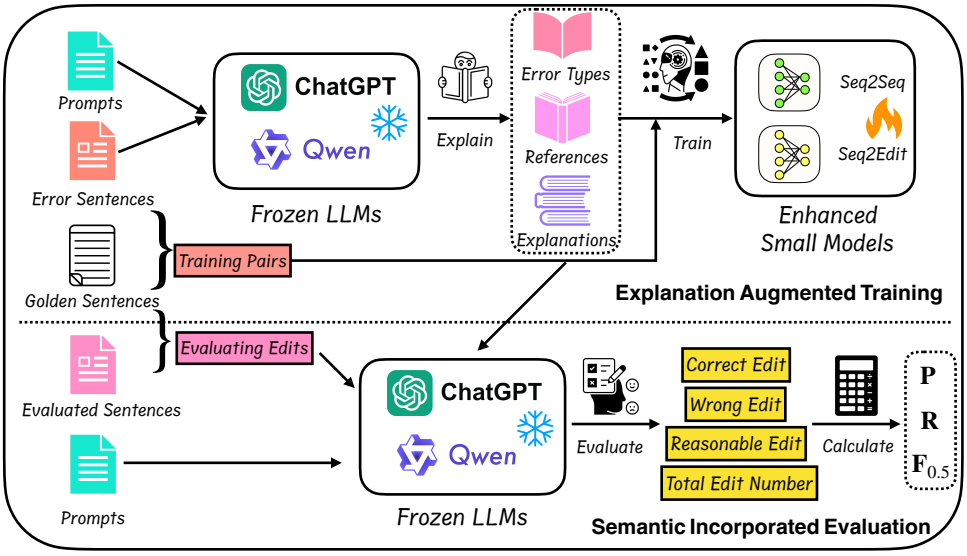

Figure 2: Our designed frameworks of EXAM and SEE.

Error Types We think that if the CGEC model knows the type of grammatical errors made in the sentence to be corrected, this will help it reduce the search scope when correcting errors, thereby helping it to make better corrections. Therefore, we ask LLMs to give the error types based on the input error sentences. Specifically, we pre-define types of common grammatical errors involving punctuation errors, spelling errors, word errors, syntax errors, etc. Then, we provide the defined error type schema along with the prompt to the LLMs, requiring them to choose only among the types we specified in the instruction prompt.

References We observe that LLMs have a particular ability to generate correct sentences based on wrong sentences, but the sentences they generate are not very controllable. Although the corrected sentences by LLMs cannot be used as the final result, we think they must be used as intermediate references for small models! Utilizing corrections from LLMs as references can provide valuable cues for the small models, thereby enhancing their performance. Therefore, we also guide LLMs to make corrections they think are reasonable for the wrong sentences and send the corrections provided by LLMs as references to the small model.

Explanations To obtain high-quality explanations from LLMs, we define three dimensions of criteria to constrain LLMs: (1) *Fluency* aims to ensure that the explanation text generated by LLMs has no grammatical errors and is fluent in expression; (2) *Rationality* requires LLMs to explain grammatical errors as humanly as possible; (3) *Comprehensiveness* is to ensure that all grammatical errors in the wrong sentences can be explained as much as possible. Additionally, we also ask LLMs to rank multiple grammatical errors in a sentence according to error severity, that is, to generate explanations for important errors first.

After LLMs explain the samples in the dataset, we concat the obtained error types, references, and explanations to the front of the original input sentences, and then send contacted text to the small CGEC models to participate in their training or inference. In summary, the design of EXAM is simple and intuitive. **LLMs and small models each perform their respective duties and give full play to their advantages.** The stored grammatical knowledge of LLMs is mined without additional fine-tuning. The small models take advantage of the alignment of supervised learning to downstream tasks with low training costs and obtain guidance from LLMs' task-related knowledge.

3.3 Semantic-incorporated Evaluation

To alleviate the dilemma that traditional CGEC evaluation cannot flexibly adapt to the subjective characteristic of CGEC because they rely entirely on dataset annotation, we design the **SE**mantic-incorporated **E**valuation framework (**SEE**) which utilizes LLMs to comprehensively evaluate CGEC by considering complex semantics.

Specifically, we first perform comparison and alignment preprocessing based on the texts of error sentences and predicted sentences to obtain the

predicted edits of the predicted text compared to the wrong sentences. We then require LLMs to evaluate each predicted edit in three dimensions based on grammatical analysis and semantic understanding of error sentences, golden sentences, and predicted sentences: (1) *Correct Edit* ($\mathbf{N}_{\mathrm{CE}}$) means that LLMs judge that the predicted edit is effective in correcting the grammatical errors of the original sentence. (2) *Wrong Edit* ($\mathbf{N}_{\mathrm{WE}}$) means that LLMs determine that the predicted edit is invalid and cannot correct grammatical errors. (3) *Reasonable Edit* ($\mathbf{N}_{\mathrm{RE}}$) refers to model edits that are not included in golden annotations, but these edits do not cause new grammatical errors and do not affect the original semantics of the sentence. Usually, this type of edit involves some intonation particles and might be incorrectly classified as an incorrect edit by traditional metrics because it is not accounted for in the dataset annotations. From these three dimensions we design, we can know that **different from traditional evaluation indicators, LLMs do not need precise text matching to determine whether the predicted edit exists in the golden edit set to further determine whether this predicted edit is valid. The judgment of LLMs is more flexible and takes into account the semantics of the text more comprehensively.** In addition, it is worth mentioning that to make LLMs' judgment on edits more accurate, we also input the explanation information obtained in EXAM into LLMs at the same time when SEE evaluates.

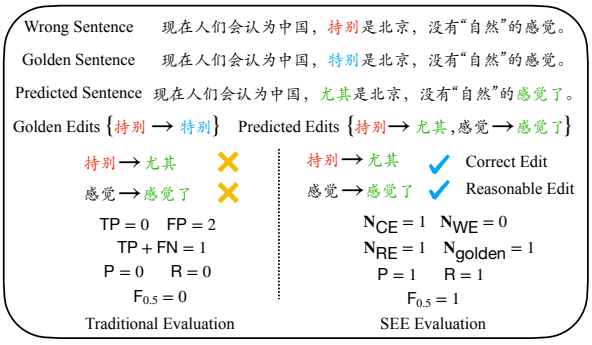

Figure 3: The comparison examples of evaluation.

Based on the above three values derived from LLMs, we can calculate Precision, Recall, and $\mathrm{F}_{0.5}$ scores as follows:

$$\mathrm{P} = \frac{\mathbf{N}_{\mathrm{CE}}}{\mathbf{N}_{\mathrm{CE}} + \mathbf{N}_{\mathrm{WE}}}, \tag{1}$$

$$\mathrm{R} = \frac{\mathbf{N}_{\mathrm{CE}}}{\mathbf{N}_{\mathrm{golden}}}, \tag{2}$$

$$\mathrm{F}_{0.5} = \frac{(1+0.5^2) \times \mathrm{P} \times \mathrm{R}}{0.5^2 \times \mathrm{P} + \mathrm{R}}, \tag{3}$$

where $\mathbf{N}_{\mathrm{golden}}$ is the length of the golden edit set for the wrong sentence. The $\mathrm{F}_{0.5}$ score is widely used in GEC-related studies because GEC is an application that pays more attention to precision. Furthermore, to better explain the mechanism of SEE, we provide an evaluation example in Figure 3.

To enable LLMs to perform the tasks we design for EXAM and SEE, while we input prompts into LLMs, we also input task demonstration examples to LLMs to make them follow our instructions more through in-cotext learning. Due to the limitation of pages, the specific contents of our designed prompts for instructing LLMs to accomplish corresponding goals are presented in Appendix B.

4 Experiments

4.1 Experiment Setup

Datasets We mainly use the HSK dataset (Zhang, 2009) as training data. In our experiments, there are two settings for the use of training data: (1) **Full HSK data**, that is, using all 156,870 samples for model training; (2) **Sampled HSK data**, we randomly sample approximately 10% of the HSK data, that is, 15,000 samples for model training. In terms of test data, the CGEC data can be divided into two types of test data according to the source of the grammatical error sentences, namely Chinese-as-Second-Language (CSL) and Chinese native speaker data. To ensure the breadth of our experiment, we select the **NLPCC test data** (Zhao et al., 2018) which is the CSL data, and the **NaCGEC benchmark** (Ma et al., 2022) which is Chinese native speaker data as the test sets of our experiment. The NLPCC test data contains 2,000 samples and NaCGEC contains 5,869 wrong sentences.

Evaluation Metrics To ensure the comparability of our experiments with previous CGEC works, in addition to using our own designed **SEE** to evaluate P/R/$\mathrm{F}_{0.5}$, we also report the widely used traditional **word/character-level** P/R/$\mathrm{F}_{0.5}$. Particularly, as in the previous work (Zhang et al., 2022), we also apply the MaxMatch scorer (Dahlmeier and Ng, 2012) and PKUNLP word segmentation tool (Zhao et al., 2018) to obtain the word-level performance. Therefore, to verify the effectiveness of our designed EXAM, we also conduct **human evaluation** experiments to provide the real performance of the models from a human perspective.

Training Data	Model	Word-Level			Character-Level			SEE		
		P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$
None	`GPT-3.5-Turbo`	24.36	28.01	25.01	27.71	29.19	27.99	53.82	30.14	46.51
None	`Qwen-72B-Chat`	**27.88**	**32.85**	**28.75**	**32.42**	**34.97**	**32.90**	**67.20**	**35.01**	**56.76**
Sampled (15K)	mT5-Base	16.10	8.93	13.87	30.25	8.77	20.30	58.36	9.89	29.47
Full (156K)	mT5-Base	24.08	16.74	22.14	38.37	17.14	30.75	67.37	19.37	45.05
Sampled (15K)	w/ EXAM (GPT)	25.21$^{\uparrow}$	17.76$^{\uparrow}$	23.26$^{\uparrow}$	**39.04**$^{\uparrow}$	18.16$^{\uparrow}$	31.74$^{\uparrow}$	69.29$^{\uparrow}$	20.27$^{\uparrow}$	46.70$^{\uparrow}$
Sampled (15K)	w/ EXAM (Qwen)	**26.41**$^{\uparrow}$	**20.57**$^{\uparrow}$	**25.00**$^{\uparrow}$	38.76$^{\uparrow}$	**21.81**$^{\uparrow}$	**33.55**$^{\uparrow}$	**69.76**$^{\uparrow}$	**22.63**$^{\uparrow}$	**49.25**$^{\uparrow}$
Sampled (15K)	BART-Large	19.46	14.77	18.30	32.07	13.67	25.27	62.94	12.18	34.33
Full (156K)	BART-Large	**28.35**	22.30	26.89	39.10	22.75	34.19	63.16	17.31	41.29
Sampled (15K)	w/ EXAM (GPT)	28.33$^{\uparrow}$	**23.38**$^{\uparrow}$	**27.17**$^{\uparrow}$	39.61$^{\uparrow}$	**23.87**$^{\uparrow}$	**35.00**$^{\uparrow}$	**68.55**$^{\uparrow}$	**23.31**$^{\uparrow}$	**49.38**$^{\uparrow}$
Sampled (15K)	w/ EXAM (Qwen)	27.91$^{\uparrow}$	22.24$^{\uparrow}$	26.55$^{\uparrow}$	**40.01**$^{\uparrow}$	22.90$^{\uparrow}$	34.81$^{\uparrow}$	62.94$^{\uparrow}$	22.18$^{\uparrow}$	46.02$^{\uparrow}$
Sampled (15K)	GECToR-Chinese	10.85	6.40	9.53	**29.49**	4.65	14.26	55.60	4.41	16.74
Full (156K)	GECToR-Chinese	**18.26**	10.99	16.12	27.03	11.99	21.60	48.32	**12.21**	30.36
Sampled (15K)	w/ EXAM (GPT)	18.09$^{\uparrow}$	**12.74**$^{\uparrow}$	**16.69**$^{\uparrow}$	27.53$^{\uparrow}$	**12.71**$^{\uparrow}$	**22.32**$^{\uparrow}$	**49.46**$^{\uparrow}$	12.05$^{\uparrow}$	**30.51**$^{\uparrow}$
Sampled (15K)	w/ EXAM (Qwen)	17.31$^{\uparrow}$	12.06$^{\uparrow}$	15.92$^{\uparrow}$	25.95$^{\uparrow}$	11.63$^{\uparrow}$	20.82$^{\uparrow}$	48.98$^{\uparrow}$	11.49$^{\uparrow}$	29.63$^{\uparrow}$

Table 1: Performance of various models on the NLPCC test set. Note that 15K and 156K represent the amount of HSK data. $^{\uparrow}$ means that EXAM has improved performance compared to the baselines with the same training data.

Baselines and Base Models The current mainstream CGEC models are mainly divided into two categories, namely Seq2Seq and Seq2Edit models. Since our EXAM framework is model-agnostic, we select the **representative Seq2Seq and Seq2Edit** models as baselines: (1) **BART-Large** (Katsumata and Komachi, 2020) and **mT5-Base** (Xue et al., 2021) are Seq2Seq models for text generation and can be straightforwardly trained for CGEC; (2) **GECToR-Chinese** (Omelianchuk et al., 2020) is the most widely used **Seq2Edit** method for CGEC. In addition, we select `GPT-3.5-Turbo` (OpenAI, 2023) and `Qwen-72B-Chat` (Alibaba, 2023) as the explainer-LLMs respectively. As for the evaluator-LLMs in SEE, we recommend the most advanced `GPT-4-Turbo` (OpenAI, 2023).

Implementation Details We utilize Chinese-BART-Large (Shao et al., 2021), Mengzi-T5-Base (Chinese) (Zhang et al., 2021), Chinese-Struct-Bert-Large (Wang et al., 2020) to initialize small models. For open-source LLMs, we run their inference process on 4 NVIDIA A100 GPUs. For closed-source LLMs, we directly access them through the official APIs. It is worth noting that in all our reported experiments, EXAM provides only one error type/reference/explanation information for each incorrect sentence. Because our experiments are only verification experiments, for better performance, researchers can obtain more explanation information to enhance the small models in EXAM. The specific prompts used by our method are in Appendix B, and other implementation details and hyperparameter selection are in Appendix A.

4.2 Main Results

Our main results on NLPCC are presented in Table 1, we also provide main results and analyses on NaCGEC in Appendix C and Table 6.

Main Results of EXAM From Table 1, we can know that: (1) With the same amount of training data, EXAM generally brings significant improvements to all baselines under all evaluation metrics. (2) With only 10% of the labeled training data, small models enhanced by EXAM achieve performance equivalent to or better than that of training with the full amount of data. (3) The model-agnostic nature of EXAM enables it to bring stable gains no matter what LLMs are selected, or for small models of Large/Base scale.

Main Results of SEE From Table 1, we see that: (1) The evaluation results of SEE are basically consistent in trend with traditional metrics, which shows the correctness of SEE. (2) Especially for the results of LLMs, we observe that SEE achieves a huge numerical difference from the results obtained by traditional metrics, which indicates that SEE is more suitable for GEC evaluation in the era of LLMs. Note that the base model of SEE is `GPT-4-Turbo`, which is different from the evaluated LLMs, so it will not cause unfair evaluation.

4.3 Analyses and Discussion

4.3.1 The Impact of Fine-grained Explanation Information on EXAM

The main results of EXAM are obtained jointly from three kinds of information error

types/references/explanations from LLMs, so it is necessary for us to conduct ablation studies on the three kinds of information to observe their respective contributions to EXAM. As shown in Table 2, we conduct ablation experiments on NLPCC test data with `GPT-3.5-Turbo` as the base model of EXAM and BART-Large as the enhanced small model. We can see that each type of information can bring significant improvements to BART-Large when executed individually, demonstrating the correctness of our choice of obtaining information from LLMs. In particular, the references have the greatest improvement for the small model, which shows that the correction results made by LLMs can bring good reference and guidance to the small model, and a good reference correction result can bring the most direct gain to the small model. Furthermore, we see that when various types of information are used in pairs, performance can be further improved compared to individual information. This shows that the compatibility between the three types of information we designed is very good and would not affect each other.

Method	Word-$F_{0.5}$	Char-$F_{0.5}$
BART-Large	18.30	25.27
+ Error Types	21.74$^{\uparrow}$	29.12$^{\uparrow}$
+ References	23.88$^{\uparrow}$	33.49$^{\uparrow}$
+ Explanations	21.52$^{\uparrow}$	29.84
+ Error Types + References	24.21$^{\uparrow}$	33.66$^{\uparrow}$
+ Error Types + Explanations	23.29$^{\uparrow}$	32.54$^{\uparrow}$
+ References + Explanations	25.18$^{\uparrow}$	33.74$^{\uparrow}$
BART-Large w/ EXAM (GPT)	**27.17**	**35.00**

Table 2: Ablation results for fine-grained explanation information. The training data for all models is 15K sampled HSK data. The test data is NLPCC. Note that the BART-Large w/ EXAM (GPT) is equivalent to BART-Large+Error Types+References+Explanations.

Method	Word-$F_{0.5}$	Char-$F_{0.5}$
BART-Large	18.30	25.27
Train (No gold) / Test (No gold)	27.17^{-}	35.00^{-}
Train (Gold) / Test (No gold)	21.57$^{\downarrow}$	28.93$^{\downarrow}$
Train (No gold) / Test (Gold)	25.98$^{\downarrow}$	37.56$^{\uparrow}$
Train (Gold) / Test (Gold)	43.10$^{\uparrow}$	60.40$^{\uparrow}$
BART-Large w/ EXAM (GPT)	27.17	35.00

Table 3: The impact of golden annotation information. The training data is 15K sampled HSK data. The test data is NLPCC. Note that the BART-Large w/ EXAM (GPT) is equivalent to Train (No gold) / Test (No gold).

4.3.2 The Impact of Golden Annotation Information on EXAM

To further explore the performance upperbound of EXAM, in the process of using LLMs to obtain the training data and test data of the small model, we input the golden sentences annotated by the dataset into the LLMs to observe the performance changes of the small model. In other words, we want to observe how the quality of the explanation information generated by LLMs changes when it accepts golden sentences as input. In Table 3, we are surprised to find that when we add golden sentences in the process of LLMs generating training data or generating test data, the model performance declines compared to not adding golden sentences in both processes (i.e., Train (No gold)/ Test (No gold)). This is an interesting and counter-intuitive phenomenon, and we think it shows the difference and gap between the generative paradigm of LLMs and the golden sentences annotated in the dataset. If LLMs are only allowed to see golden sentences during training or testing, this will cause the explanation information generated by LLMs to be very different from what it tends to generate on its own, resulting in a gap between the training and test data of the small model, which leads to performance degradation of small models. Therefore, we can also understand why there is a huge performance gain when inputting golden sentences to LLMs in both training and testing processes. In this case, LLMs generate sentences similar to golden sentences in both training data and test data.

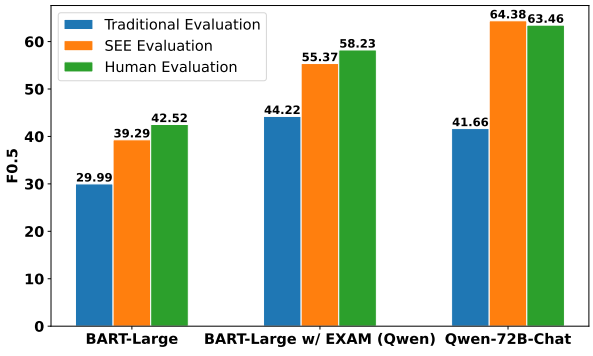

Figure 4: Human evaluation results. The training data is 15K sampled HSK data. The test data is 200 sampled NLPCC data. The traditional metric is Char-$F_{0.5}$.

4.3.3 Human Evaluation for SEE

The design motivation of SEE is to use LLMs to bring evaluation more consistent with the human perspective to CGEC. Therefore, we conduct hu-

Error Sentence	这段话给我们有道理的，虽然现在黑暗，但等着、忍着，光明会到来的。
Golden Sentence	这段话是有道理的，虽然现在黑暗，但等着、忍着，光明就会到来。 This paragraph is reasonable. Although it is dark now, if we wait and endure, the light will come.
Error Type **GPT-3.5-Turbo**	标点误用，句式杂糅 Misuse of punctuation, mixed sentence structures
Reference **GPT-3.5-Turbo**	这段话给我们的道理：虽然现在黑暗，但等着、忍着，光明会到来的。
Explanation **GPT-3.5-Turbo**	在这个句子中，动词“给我们”和“有道理”成分赘余，此处改为“给我们的道理”，并且下句为道理的内容，因此不应该使用“，”而应该使用表示承接的“：”。
Error Type **Qwen-72B-Chat**	词语冗余, 逻辑不通 Redundant words, illogical logic
Reference **Qwen-72B-Chat**	这段话给我们提供了有道理的启示，虽然现在处于黑暗时期，但只要坚持等待、忍受，光明总会到来的。
Explanation **Qwen-72B-Chat**	“给我们”与“有道理的”都表示“这段话”的意义，造成了词语冗余。应将“给我们”改为“提供了”。接着，“虽然现在黑暗，但等着、忍着，光明会到来的”缺乏逻辑连贯性，应补充关联词语“虽然”。

Table 4: We mark changes made by LLMs in orange and poor explanations given by LLMs in red.

man evaluation experiments to observe whether SEE or traditional metrics are closer to human. Specifically, we randomly select 200 test samples from NLPCC, then require three annotators to judge the correction results of models separately, and calculate the average P/R/$F_{0.5}$ scores of human evaluation based on the three annotators' judgment results. From Figure 4, we see that: (1) For various models, SEE's evaluation is closer to human evaluation than traditional evaluation, which shows that our designed SEE can more realistically measure the CGEC performance than traditional evaluation. (2) SEE's evaluation of LLMs differs very little from human evaluation, indicating that SEE is more suitable for the evaluation of LLMs. (3) Unlike the cases where evaluation results for small models fall below human evaluation, SEE's evaluation of LLMs can slightly surpasses human evaluation results. This is because SEE relies on another LLM (i.e., `GPT-4-Turbo`) for its evaluation process, indicating better understanding among LLMs.

4.4 Case Observation

To verify the correctness of our motivation for using LLMs as explainers, and to demonstrate the explanation information generated by EXAM, we give cases in Table 4 of `GPT-3.5-Turbo` and `Qwen-72B-Chat` acting as the explainer respectively. We can see from Table 4 that, although the two LLMs make different error-type judgments, they both give their own reasonable explanations for their error-type judgments. Regarding the reference corrections they give, we see that `Qwen-72B-Chat` prefers free generation compared to `GPT-3.5-Turbo`. Of course, we think the corrected sentence generated by `Qwen-72B-Chat` is more fluent and reasonable. For the explanations of grammatical errors made in the wrong sentence, we can see that both LLMs give quality explanations to a certain extent. Although there are some minor flaws, on the whole, they can give explanations that can be helpful for humans or small models to be enhanced. Additionally, we also provide more cases in which LLMs do explanations and evaluations in the form of data supplementary materials.

5 Conclusion

In this paper, focusing on the dilemma that LLMs cannot achieve satisfactory results as correctors on CGEC, we rethink how LLMs should be effectively utilized in the CGEC task. To fully exploit the rich grammatical knowledge and powerful semantic understanding ability of LLMs, and bypass the main reason why the LLMs corrector is not suitable for the CGEC task, that is, the minimum change principle, we propose the training framework EXAM that uses LLMs as explainers to enhance CGEC small models, and the novel evaluation method SEE that utilizes LLMs as evaluators to give more reasonable evaluation of the CGEC task. Extensive empirical results and analyses show that our work is a meaningful exploration of how LLMs and small models can coexist and make progress together on downstream tasks such as CGEC.

Limitations

Currently, the main limitation of our work is the scope of the languages. As we all know, GEC in various languages has its application significance, so it is valuable to apply our methods to other languages further. The main reason why we did not apply our methods to languages such as English is that there are many differences in the types of grammatical errors and grammatical rules that CGEC and EGEC focus on. Therefore, the prompts of EXAM and SEE need to be re-customized when applied to the English scenario. The purpose of our paper is to rethink how LLMs should be appropriately utilized in the GEC field. Changing prompts to adapt to new languages is not the main technical contribution and innovation we pursue. In the future, to enhance the impact of our work and serve a wider community, we will expand EXAM and SEE to the English scenario.

Ethics Statement

The data and models (including LLMs) used in our experiments are all publicly available academic resources. We also paid for closed-source LLMs that require charging for APIs, so there is no ethical issue about data or models in our work.

References

Alibaba. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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.

Christopher Davis, Andrew Caines, Øistein E. 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. *CoRR*, abs/2401.07702.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2023. Grammargpt: Exploring open-source llms for native chinese grammatical error correction with supervised fine-tuning. In *Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part III*, volume 14304 of *Lecture Notes in Computer Science*, pages 69–80. Springer.

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. *CoRR*, abs/2304.01746.

Masahiro Kaneko and Naoaki Okazaki. 2023a. Controlled generation with prompt insertion for natural language explanations in grammatical error correction. *CoRR*, abs/2309.11439.

Masahiro Kaneko and Naoaki Okazaki. 2023b. Reducing sequence length by predicting edit spans with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10017–10029. Association for Computational Linguistics.

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.

Sang Yun Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Beyond english: Evaluating llms for arabic grammatical error correction. In *Proceedings of ArabicNLP 2023, Singapore (Hybrid), December 7, 2023*, pages 101–119. Association for Computational Linguistics.

Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2023a. Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce. *CoRR*, abs/2308.06966.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, and Qingyu Zhou. 2023b. On the (in) effectiveness of large language models for chinese text correction. *arXiv preprint arXiv:2307.09007*.

Aiwei Liu, Xuming Hu, Lijie Wen, and Philip S. Yu. 2023. A comprehensive evaluation of chatgpt's zero-shot text-to-sql capability. *CoRR*, abs/2303.13547.

Shirong Ma, Yinghui Li, Rongyi Sun, Qingyu Zhou, Shulin Huang, Ding Zhang, Yangning Li, Ruiyang Liu, Zhongli Li, Yunbo Cao, Haitao Zheng, and Ying Shen. 2022. Linguistic rules-based corpus generation for native chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 576–589. 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.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Robert Östling, Katarina Gillholm, Murathan Kurfali, Marie Mattson, and Mats Wirén. 2023. Evaluation of really good grammatical error correction. *CoRR*, abs/2308.08982.

Maria Carolina Penteado and Fábio Perez. 2023. Evaluating GPT-3.5 and GPT-4 on grammatical error correction for brazilian portuguese. *CoRR*, abs/2306.15788.

Fanyi Qu and Yunfang Wu. 2023. Evaluating the capability of large-scale language models on chinese grammatical error correction task. *CoRR*, abs/2307.03972.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. CPT: A pre-trained unbalanced transformer for both chinese language understanding and generation. *CoRR*, abs/2109.05729.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2023. Gee! grammar error explanation with large language models. *CoRR*, abs/2311.09517.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of GPT-4: reliably evaluating large language models on sequence to sequence tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 8776–8788. Association for Computational Linguistics.

Chang Su, Xiaofeng Zhao, Xiaosong Qiao, Min Zhang, Hao Yang, Junhao Zhu, Ming Zhu, and Wenbing Ma. 2023. Hwcgec:hw-tsc's 2023 submission for the nlpcc2023's chinese grammatical error correction task. In *Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part III*, volume 14304 of *Lecture Notes in Computer Science*, pages 59–68. Springer.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Jiangnan Xia, Zuyi Bao, Liwei Peng, and Luo Si. 2020. Structbert: Incorporating language structures into pre-training for deep language understanding. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Jingheng Ye, Yinghui Li, Qingyu Zhou, Yangning Li, Shirong Ma, Hai-Tao Zheng, and Ying Shen. 2023. CLEME: Debiasing multi-reference evaluation for grammatical error correction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6174–6189, Singapore. Association for Computational Linguistics.

Tianyu Yu, Chengyue Jiang, Chao Lou, Shen Huang, Xiaobin Wang, Wei Liu, Jiong Cai, Yangning Li, Yinghui Li, Kewei Tu, et al. 2023. Seqgpt: An out-of-the-box large language model for open domain sequence understanding. *arXiv preprint arXiv:2308.10529*.

Baolin Zhang. 2009. Features and functions of the hsk dynamic composition corpus. *International Chinese Language Education*, 4:71–79.

Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. 2023. Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance. *CoRR*, abs/2305.13225.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022. Mucgec: a multi-reference multi-source evaluation dataset for chinese grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3118–3130. Association for Computational Linguistics.

Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *CoRR*, abs/2110.06696.

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 - 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26-30, 2018, Proceedings, Part II*, volume 11109 of *Lecture Notes in Computer Science*, pages 439–445. Springer.

A Implementation Details and Hyperparameters

The hyperparameter values of the small models to be enhanced in our experiments are shown in Table 5. Besides, the loss functions for Seq2Seq models are the label-smoothed cross-entropy, and the loss function for Seq2Edit is cross-entropy.

B Our Designed Prompts for EXAM and SEE

In order to guide LLMs to achieve our designed tasks as we expect, we carefully design the instruction prompts based on the characteristics of the CGEC task. The prompts for explanation are as shown in Figure 5, and the prompts for evaluation are as shown in Figure 6.

In addition, as mentioned in the main text of this paper, to make the results generated by LLMs more accurate, we also input task examples (or demonstrations) to LLMs to stimulate their In-context Learning capabilities. Considering that the prompts with in-context learning examples added are very long, we upload the prompts with task examples in the form of software supplementary materials to facilitate peer review.

C Main Results on NaCGEC

The main results of EXAM and SEE on NaCGEC are presented in Table 6. Note that the models we test on NaCGEC are all trained using HSK data. The HSK data comes from sentences with grammatical errors made by foreigners when learning Chinese, while NaCGEC comes from the grammatical errors made by native Chinese speakers in daily life. Ma et al. have proven that Chinese native CGEC data such as NaCGEC is more difficult than CSL data such as HSK because the grammatical errors made by native speakers are more subtle than those made by foreigners. Therefore, as shown in Table 6, when CGEC models trained with HSK data are tested on NaCGEC, low performance is understandable and expected.

From Table 6, we can get similar conclusions as on NLPCC. EXAM can bring stable and competitive enhancements to small models with the participation of small-scale training data, and the performance enhanced by EXAM is comparable to the performance of small models trained with full-scale data. Meanwhile, SEE can still bring reliable evaluation to CGEC models. The experiment on NaCGEC reflects the robustness of our proposed EXAM and SEE to different data sources, that is, they are effective for both CSL CGEC data and native CGEC data.

Configurations	BART-Large	mT5-Base	GECToR-Chinese
Model type	Seq2Seq	Seq2Seq	Seq2Edit
Epochs	10	10	20 (2 cold epochs)
Batch size	256	256	128
Optimizer	Adam	Adam	Adam
β_1	0.9	0.9	0.9
β_2	0.999	0.999	0.999
ϵ	1×10^{-8}	1×10^{-8}	1×10^{-8}
Learning rate	3×10^{-6}	5×10^{-5}	1×10^{-5}($1 \times 10^{-3}$for cold)

Table 5: Hyperparameter values of the small models to be enhanced in our experiments.

Training Data	Model	Word-Level			Character-Level			SEE		
		P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$
None	`GPT-3.5-Turbo`	13.84	**11.67**	13.35	9.58	**9.66**	9.59	**39.65**	**12.17**	**27.31**z
None	`Qwen-72B-Chat`	**14.23**	11.33	**13.53**	**10.32**	8.83	**9.98**	32.55	4.74	23.14
Sampled (15K)	mT5-Base	5.38	0.65	2.19	4.5	0.64	2.03	**36.11**	4.40	14.79
Full (156K)	mT5-Base	2.78	3.72	2.93	1.98	3.17	2.14	18.25	8.20	14.65
Sampled (15K)	w/ EXAM (GPT)	**11.06**$^{\uparrow}$	**4.03**$^{\uparrow}$	**8.20**$^{\uparrow}$	**8.34**$^{\uparrow}$	**3.51**$^{\uparrow}$	**6.54**$^{\uparrow}$	34.26$^{\downarrow}$	**8.80**$^{\uparrow}$	**21.70**$^{\uparrow}$
Sampled (15K)	w/ EXAM (Qwen)	10.51$^{\uparrow}$	3.11$^{\uparrow}$	7.12$^{\uparrow}$	7.60$^{\uparrow}$	2.55$^{\uparrow}$	5.44$^{\uparrow}$	32.66$^{\downarrow}$	7.70$^{\uparrow}$	19.81$^{\uparrow}$
Sampled (15K)	BART-Large	7.07	2.34	5.04	5.59	2.15	4.24	29.45	5.96	16.46
Full (156K)	BART-Large	**11.08**	4.07	**8.24**	**9.39**	4.05	**7.43**	**39.34**	9.01	**23.52**
Sampled (15K)	w/ EXAM (GPT)	10.11$^{\uparrow}$	**4.48**$^{\uparrow}$	8.08$^{\uparrow}$	8.64$^{\uparrow}$	**4.49**$^{\uparrow}$	7.29$^{\uparrow}$	30.00$^{\uparrow}$	**9.50**$^{\uparrow}$	20.97$^{\uparrow}$
Sampled (15K)	w/ EXAM (Qwen)	8.46$^{\uparrow}$	3.52$^{\uparrow}$	6.60$^{\uparrow}$	7.06$^{\uparrow}$	3.41$^{\uparrow}$	5.81$^{\uparrow}$	31.22$^{\uparrow}$	5.99$^{\uparrow}$	16.94$^{\uparrow}$
Sampled (15K)	GECToR-Chinese	2.40	0.11	0.46	3.82	0.19	0.80	26.31	3.08	10.48
Full (156K)	GECToR-Chinese	8.53	1.12	3.67	4.22	0.93	2.47	27.89	3.23	11.03
Sampled (15K)	w/ EXAM (GPT)	**12.08**$^{\uparrow}$	2.19$^{\uparrow}$	6.35$^{\uparrow}$	**9.26**$^{\uparrow}$	1.87$^{\uparrow}$	5.17$^{\uparrow}$	30.55$^{\uparrow}$	4.74$^{\uparrow}$	14.62$^{\uparrow}$
Sampled (15K)	w/ EXAM (Qwen)	11.09$^{\uparrow}$	**2.63**$^{\uparrow}$	**6.74**$^{\uparrow}$	9.01$^{\uparrow}$	**1.96**$^{\uparrow}$	**5.24**$^{\uparrow}$	**31.35**$^{\uparrow}$	**5.01**$^{\uparrow}$	**15.28**$^{\uparrow}$

Table 6: Performance of various models on the NaCGEC benchmark. Note that 15K and 156K represent the amount of HSK data. $^{\uparrow}$ means that EXAM has improved performance compared to the baselines with the same training data.

你是一个优秀的语法纠错解释模型，能针对中文文本中的标点错误、拼写错误、词语错误和句法错误等提供流畅、合理且忠实的解释。

你需要识别我输入的句子中可能含有的语法错误并纠正句子，对错误句中的标点错误、拼写错误、词语错误和句法错误等提供流畅、合理且忠实的解释，解释包括语法错误类型和解释描述信息。流畅性要求解释本身没有语法错误且表达流畅；合理性要求对语法错误的解释是能被人们接受的；忠实性要求对句子中所有语法错误都有对应解释，且解释能对应正确句的纠正方式。为了提升解释的合理性和忠实性，你需要：
1） 提供充分且全面的纠正证据词。
2） 必须根据纠正句给出合理的语法规则。最好使用三段论推理方式给出解释。
3） 如果一处编辑改动(edit)存在多个语法错误，请按照优先级顺序：句法级别错误>词语级别错误>拼写级别错误>标点级别错误，选择优先级最高的语法错误进行解释。
4） 每个编辑改动(edit)分别给出相应的严重程度、错误类型和解释描述。
5） 错误类型"error_type"只能是以下二级错误类型，即：
    a） 标点冗余、标点丢失、标点误用；
    b） 字音混淆错误、字形混淆错误、词内部字符异位错误、命名实体拼写错误；
    c） 词语冗余、词语丢失、词语误用
    d） 词序不当、逻辑不通、句式杂糅
    e） 照应错误、歧义错误、语气不协调
中的一个或者多个，不能单纯写“句子级错误”,“词级错误”以及“标点级别错误”。
6） 当不能确定是那个错误类型时，统一写为“其他句子级错误”或者“其他词级错误”。

请注意你需要强调解释描述信息中的证据词和纠正方式：
- 证据词必须是出现在错误句中的文本段，并且前后使用【】包围。
- 纠正方式必须是出现在纠正句中的文本段，并且前后使用{}包围。

错误类型严格按照给出的进行解释，不可自主捏造，如果错误类型都无法匹配则标为“其他错误”。

现在开始解释：

Figure 5: Our designed explanation prompt for EXAM.

你是一个优秀的语法纠错评估模型，能针对中文文本中的标点错误、拼写错误、词语错误和句法错误等提供准确的评估。

你需要仔细对比预测句和参考句的前提上，对原错误句中的标点错误、拼写错误、词语错误和句法错误等是否被正确纠正提供合理且忠实的判断，并且对没有的被正确纠正的部分提供合理解释。

输入格式为：
```
{
    "error_sentence": 含有语法错误的句子
    "correct_sentence": 正确被语法纠正的参考句
    "edits": list 结构，包含 error_sentence 中的错误纠正信息
    "predict_sentence": 待评估的预测句，这其中只会包含对 error_sentence 的一个语法错误位置进行替换修改替换，即只替换了 error_sentence 句中的一处，你需要在 edit 中相同编辑位置的纠正进行对比判断。
}
```

输入格式为：
```
{
    "Correct Edit": bool 值，满足要求，即足够准确则为 1，否则为 0。
    "Wrong Edit": bool 值，如果 predict_sentence 中错误地修正了本来正确的部分则为 1，否则为 0。
    "Reasonable Edit": bool 值，如果不在 edit 范围附近的纠正，但是判断合理的，则为 1，否则为 0。
    "Explanation": 如果判断为不准确时，给出合理的解释，解释为什么不准确；如果准确则为"无"。
}
```
注意：输入输出都为合法的 json 格式结构

要求：
1）请仔细对比评估 predict_sentence 和 correct_sentence，并且结合语义，参考 correct_sentence，判断 predict_sentence 中的对于 error_sentence 的这一位置的语法 错误纠正是否足够准确。
2）主要关注 predict_sentence 中和 correct_sentence 词组不同的位置，首先判断是否为同一范围内语法错误，如果是 edit 范围附近没有的纠正而 predict_sentence 中有， 则 Correct Edit 是为 0，并且进一步判断是否是一个合理的纠正如果是则可 Wrong Edit 记为 1，如果判断是不合理的，则是错误地修正了本来正确的部分，Wrong Edit 要为 1；之后判断 predict_sentence 中和 correct_sentence 的纠正词是否都能准确的纠正这个语法错误。如果都能准确且合理的纠正这个错误，则输出的 Correct Edit 赋值为 1，否则为 0，并给出不准确的理由
3）Correct Edit：如果能准确且合理的纠正这个错误，则为 1，否则为 0。Wrong Edit：如果是 edit 中没有的纠正，但是是合理且准确的可以认为是合理的纠正，但如果是不合理的，则为错误地纠正，应该为 1。因此不存在 Correct Edit 和 Wrong Edit 同为 1 的情况。

现在开始进行评估：

Figure 6: Our designed evaluation prompt of SEE.