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

Yinghui Li[1]*, **Shang Qin**[1]*, **Jingheng Ye**[1], **Haojing Huang**[1], **Yangning Li**[1,2], **Shu-Yu Guo**[1]
Libo Qin[3], **Xuming Hu**[4], **Wenhao Jiang**[5]†, **Hai-Tao Zheng**[1,2]†, **Philip S. Yu**[6]
[1] Shenzhen International Graduate School, Tsinghua University, [2] Peng Cheng Laboratory
[3] School of Computer Science and Engineering, Central South University
[4] The Hong Kong University of Science and Technology (Guangzhou)
[5] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[6] University of Illinois Chicago

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 the challenging nature of the task. 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 to the CGEC small models during error correction, aiming 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 topic in the NLP community. In the vast Chinese NLP research field, 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 CGEC performance of LLMs 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.

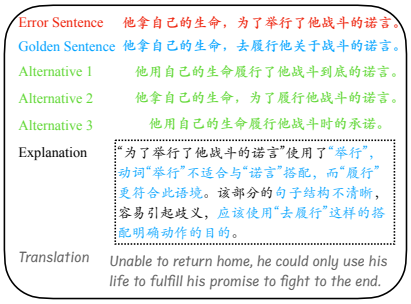

Figure 1: The example of subjectivity and explainability of CGEC. The explanation is produced by ChatGPT.

To address this dilemma, our work rethinks the proper utilization of LLMs to promote the development of CGEC. Overviewing recent research trends, the subjectivity and explainability of GEC have received great attention (Ye et al., 2023c; Song et al., 2023; Kaneko and Okazaki, 2023a). As illustrated in Figure 1, a grammatically incorrect sentence often has different correction 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 incorrect sentence contains instructive information and knowledge for error correction. If we can obtain high-quality explanations of incorrect sentences, it will

*indicates equal contribution. E-mails: {liyinghu20, qin-s23}@mails.tsinghua.edu.cn
† Corresponding authors (cswhjiang@gmail.com, zheng.haitao@sz.tsinghua.edu.cn).
[1]Our code and used data are available at https://github.com/THUKElab/LLM4CGEC.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 553–567
July 28-30, 2025

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 related to grammatically incorrect sentences from LLMs, and then utilizes mined information to enhance the training of small models. (2) SEE requires LLMs to balance the edits annotated in the golden data with the evaluated model's edits, ensuring they do not alter the original semantics of the input sentence. This ensures more accurate and comprehensive evaluation results that consider both grammar and semantics. In summary, our contributions are in four folds:

- We propose SEE, which aims to empower the evaluation of more subjective CGEC tasks through the intervention of LLMs.
- We propose EXAM, which utilizes LLMs as explainers to enhance the training of small models. This approach enables small models not only to surpass LLMs on traditional metrics but also to demonstrate competitive performance under our proposed SEE.
- **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.

2 Motivation and Methodology

2.1 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 incorrect 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.

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 enhance the performance of small models.

LLMs as Evaluator Considering the subjective nature of the CGEC task, a sentence with gram-

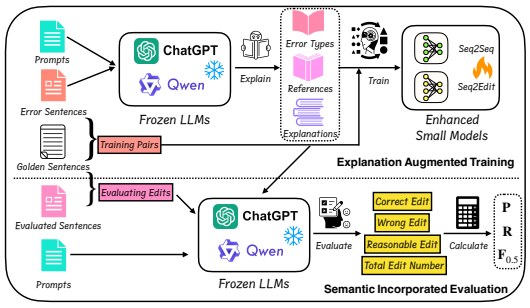

Figure 2: Our designed frameworks of EXAM and SEE.

matical 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 provides a sentence that is consistent with the original meaning of the incorrect sentence and has no grammatical errors, 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 incorrect sentences as possible. However, such an annotation process is expensive and time-consuming. **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.

2.2 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 incorrect sentences in a broad sense) we want to obtain from LLMs into three categories:

Error Types We believe that if the CGEC model knows the type of grammatical errors in the sentence to be corrected, it will help it reduce the search scope when correcting errors, thereby enabling it to make better corrections. Therefore, we ask LLMs to identify the error types based on the input sentences containing errors. 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, instructing them to choose only among the types we specified in the instruction prompt.

References We observe that LLMs have a notable ability to generate correct sentences from incorrect ones, but the sentences they produce are not highly controllable. Although the sentences corrected by LLMs cannot be used as the final result, we believe they should serve as intermediate references for small models. Using corrections from LLMs as references can provide valuable cues to the small models, thereby enhancing their performance. Therefore, we also guide LLMs to make corrections they think are reasonable for the incorrect 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 clearly and naturally as possible; (3) *Comprehensiveness* is to ensure that all grammatical errors in the incorrect 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 concatenate the obtained error types, references, and explanations to the front of the original input sentences. We then send the combined text to the small CGEC models for 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.**

2.3 Semantic-incorporated Evaluation

To address the issue that traditional CGEC evaluation cannot flexibly adapt to the subjective nature of CGEC because they rely entirely on dataset annotation, we design the **SE**mantic-incorporated **E**valuation framework (**SEE**).

Specifically, we first perform comparison and alignment preprocessing on the texts of error sentences and predicted sentences to obtain the predicted edits of the predicted text compared to the incorrect sentences. We then require LLMs to evaluate each predicted edit from three dimensions based on grammatical analysis and semantic understanding of error sentences, golden sentences, and predicted sentences: (1) *Correct Edit* ($\mathbf{N}_{\mathrm{CE}}$) indicates that LLMs judge the predicted edit to be effective in correcting the grammatical errors of the original sentence; (2) *Wrong Edit* ($\mathbf{N}_{\mathrm{WE}}$) signifies that LLMs determine that the predicted edit to be invalid and unable to correct grammatical errors; (3) *Reasonable Edit* ($\mathbf{N}_{\mathrm{RE}}$) refers to model edits not included in golden annotations, but which do not introduce 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 have designed, we can see that, **unlike different from traditional evaluation indicators, LLMs do not require precise text matching to determine whether the predicted edit exists in the golden edit set. Instead, the validity of the predicted edit is assessed more flexibly, taking into account the semantics of the text more comprehensively.**

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 incorrect 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.

To enable LLMs to perform the tasks we designed for EXAM and SEE, we input both prompts and task demonstration examples into the LLMs to facilitate their adherence to our instructions through in-context learning. Due to the limitation of pages, the specific contents of our designed prompts for instructing LLMs to accomplish corresponding goals are presented in D.1.

3 Experiments

3.1 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, 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 incorrect 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.

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` (Ope-

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	23.80	29.03	24.69	23.11	26.47	23.71	53.82	30.14	46.51
None	Qwen-72B-Chat	**27.01**	**33.87**	**28.15**	**25.87**	**29.46**	**26.52**	**67.20**	**35.01**	**56.76**
Sampled (15K)	mT5-Base	14.54	9.81	13.26	30.06	7.96	19.33	58.36	9.89	29.47
Full (156K)	mT5-Base	21.76	17.54	20.76	36.92	15.77	29.11	67.37	19.37	45.05
Sampled (15K)	w/ EXAM (GPT)	22.86$^{\uparrow}$	18.77$^{\uparrow}$	21.91$^{\uparrow}$	**37.65**$^{\uparrow}$	16.81$^{\uparrow}$	30.17$^{\uparrow}$	69.29$^{\uparrow}$	20.27$^{\uparrow}$	46.70$^{\uparrow}$
Sampled (15K)	w/ EXAM (Qwen)	**23.65**$^{\uparrow}$	**21.50**$^{\uparrow}$	**23.19**$^{\uparrow}$	36.33$^{\uparrow}$	**19.02**$^{\uparrow}$	**30.74**$^{\uparrow}$	**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**$^{\uparrow}$	21.50$^{\uparrow}$	34.13$^{\uparrow}$	62.94$^{\uparrow}$	22.18$^{\uparrow}$	46.02$^{\uparrow}$
Sampled (15K)	GECToR-Chinese	10.85	6.40	9.53	**34.89**	4.34	14.49	**55.60**	4.41	16.74
Full (156K)	GECToR-Chinese	**18.26**	10.99	16.13	27.03	11.99	21.61	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)	16.17$^{\uparrow}$	12.96$^{\uparrow}$	15.41$^{\uparrow}$	24.97$^{\uparrow}$	10.29$^{\uparrow}$	19.42$^{\uparrow}$	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.

nAI, 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).

LLMs as Correctors We selected two LLMs as Correctors to serve as baselines for comparison with our method. Specifically, we chose `Qwen-72B-Chat` and `GPT-3.5-Turbo` as our LLMs. We crafted a detailed prompt to ensure the LLMs deeply understood the task's significance when directly correcting Chinese grammatical errors (See Appendix C). Additionally, we experimented with in-context learning to enhance the performance of the LLMs. The experimental results and analysis of "LLMs as Correctors" are presented in Appendix D.

3.2 Main Results

Our main results on NLPCC are presented in Table 1, we also provide main results and analyses on NaCGEC in Appendix D.3 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.

3.3 The Impact of Fine-grained Explanation Information on EXAM

The main results of EXAM are derived from three kinds of information error types/references/explanations from LLMs. Therefore, it is necessary to conduct ablation studies on the three kinds of information to assess 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.

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.

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.

3.4 The Impact of Golden Annotation Information on EXAM

To further explore the performance upper bound of EXAM, in the process of using LLMs to obtain training and test data for 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 they are provided with 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 believe it highlights 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, the explanation information they generate will differ significantly from what they would typically produce on their own. This discrepancy can create a gap between the training and test data of the small model, leading to performance degradation. 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.

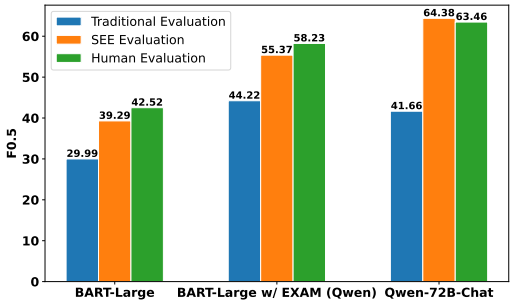

Figure 3: Human evaluation results. The training data is 15K sampled HSK data. The test data is 200 sampled NLPCC data. The traditional metric is Char-$F_{0.5}$.

3.5 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 human evaluation experiments to observe whether SEE or traditional metrics are closer to human. Specifically, we randomly select 200 test samples from NLPCC, then have three annotators to independently evaluate the models' correct results. We calculate the average P/R/$F_{0.5}$ scores of human evaluation based on the judegments from the three annotators. From Figure 3, 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, indicating better understanding among LLMs.

4 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, 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 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.

Acknowledgements

This research is supported by National Natural Science Foundation of China (Grant No. 62276154), the Natural Science Foundation of Guangdong Province (Grant No. 2023A1515012914 and 44030024103310080177O), Basic Research Fund of Shenzhen City (Grant No. JCYJ20210324120012033, JCYJ20240813112009013 and GJHZ20240218113603006), the Major Key Project of PCL (NO. PCL2024A08). This work is also supported by the Guangdong Provincial Department of Education Project (Grant No.2024KQNCX028); CAAI-Ant Group Research Fund; Scientific Research Projects for the Higher-educational Institutions (Grant No.2024312096), Education Bureau of Guangzhou Municipality; Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2025A03J3957), Education Bureau of Guangzhou Municipality. This work is also sponsored by the Excellent Young Scientists Fund in Hunan Province (2024JJ4070), the Science and Technology Innovation Program of Hunan Province under Grant 2024RC3024. This work is supported in part by NSF under grants III-2106758, and POSE-2346158.

Ethical Considerations

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.

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.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2023. A survey of natural language generation. *ACM Comput. Surv.*, 55(8):173:1–173:38.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. 2024. Llms assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5081–5099. Association for Computational Linguistics.

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.

Haojing Huang, Jingheng Ye, Qingyu Zhou, Yinghui Li, Yangning Li, Feng Zhou, and Hai-Tao Zheng. 2023a. A frustratingly easy plug-and-play detection-and-reasoning module for chinese spelling check. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 11514–11525. Association for Computational Linguistics.

Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Hai-Tao Zheng. 2023b. Lateval: An interactive llms evaluation benchmark with incomplete information from lateral thinking puzzles. *CoRR*, abs/2308.10855.

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, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Fei Huang, Jingren Zhou, and Philip S. Yu. 2025a. Benchmarking multimodal retrieval augmented generation with dynamic VQA dataset and self-adaptive planning agent. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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, Jiayi Kuang, Yangning Li, Shu-Yu Guo, Chao Qu, Xiaoyu Tan, Hai-Tao Zheng, Ying Shen, and Philip S. Yu. 2025b. Refine knowledge of large language models via adaptive contrastive learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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*.

Yinghui Li, Jiayi Kuang, Haojing Huang, Zhikun Xu, Xinnian Liang, Yi Yu, Wenlian Lu, Yangning Li, Xiaoyu Tan, Chao Qu, Ying Shen, Hai-Tao Zheng, and Philip S. Yu. 2025c. One example shown, many concepts known! counterexample-driven conceptual reasoning in mathematical llms. *CoRR*, abs/2502.10454.

Yinghui Li, Shirong Ma, Shaoshen Chen, Haojing Huang, Shulin Huang, Yangning Li, Hai-Tao Zheng, and Ying Shen. 2025d. Correct like humans: Progressive learning framework for chinese text error correction. *Expert Syst. Appl.*, 265:126039.

Yinghui Li, Shirong Ma, Qingyu Zhou, Zhongli Li, Yangning Li, Shulin Huang, Ruiyang Liu, Chao Li, Yunbo Cao, and Haitao Zheng. 2022a. Learning from the dictionary: Heterogeneous knowledge guided fine-tuning for chinese spell checking. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 238–249. Association for Computational Linguistics.

Yinghui Li, Zishan Xu, Shaoshen Chen, Haojing Huang, Yangning Li, Yong Jiang, Zhongli Li, Qingyu Zhou, Hai-Tao Zheng, and Ying Shen. 2023c. Towards real-world writing assistance: A chinese character checking benchmark with faked and misspelled characters. *CoRR*, abs/2311.11268.

Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng. 2022b. The past mistake is the future wisdom: Error-driven contrastive probability optimization for chinese spell checking. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3202–3213. Association for Computational Linguistics.

Yinghui Li, Qingyu Zhou, Yuanzhen Luo, Shirong Ma, Yangning Li, Hai-Tao Zheng, Xuming Hu, and Philip S. Yu. 2024. When llms meet cunning texts: A fallacy understanding benchmark for large language models. In *Neural Information Processing Systems*.

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.

Ruiyang Liu, Yinghui Li, Linmi Tao, Dun Liang, and Hai-Tao Zheng. 2022. Are we ready for a new paradigm shift? A survey on visual deep MLP. *Patterns*, 3(7):100520.

Shirong Ma, Yinghui Li, Haojing Huang, Shulin Huang, Yangning Li, Hai-Tao Zheng, and Ying Shen. 2023. Progressive multi-task learning framework for chinese text error correction. *CoRR*, abs/2306.17447.

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.

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.

Zhikun Xu, Yinghui Li, Ruixue Ding, Xinyu Wang, Boli Chen, Yong Jiang, Haitao Zheng, Wenlian Lu, Pengjun Xie, and Fei Huang. 2025. Let llms take on the latest challenges! A chinese dynamic question answering benchmark. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 10435–10448. Association for Computational Linguistics.

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, Yangning Li, and Hai-Tao Zheng. 2023a. Mixedit: Revisiting data augmentation and beyond for grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10161–10175. Association for Computational Linguistics.

Jingheng Ye, Yinghui Li, Shirong Ma, Rui Xie, Wei Wu, and Hai-Tao Zheng. 2022. Focus is what you need for chinese grammatical error correction. *CoRR*, abs/2210.12692.

Jingheng Ye, Yinghui Li, and Haitao Zheng. 2023b. System report for ccl23-eval task 7: Thu kelab (sz) - exploring data augmentation and denoising for chinese grammatical error correction. In *China National Conference on Chinese Computational Linguistics*.

Jingheng Ye, Yinghui Li, Qingyu Zhou, Yangning Li, Shirong Ma, Hai-Tao Zheng, and Ying Shen. 2023c. 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.

Jingheng Ye, Shang Qin, Yinghui Li, Xuxin Cheng, Libo Qin, Hai-Tao Zheng, Peng Xing, Zishan Xu, Guo Cheng, and Zhao Wei. 2024a. EXCGEC: A benchmark of edit-wise explainable chinese grammatical error correction. *CoRR*, abs/2407.00924.

Jingheng Ye, Shang Qin, Yinghui Li, Hai-Tao Zheng, Shen Wang, and Qingsong Wen. 2025. Corrections meet explanations: A unified framework for explainable grammatical error correction. *CoRR*, abs/2502.15261.

Jingheng Ye, Zishan Xu, Yinghui Li, Xuxin Cheng, Linlin Song, Qingyu Zhou, Hai-Tao Zheng, Ying Shen, and Xin Su. 2024b. CLEME2.0: towards more interpretable evaluation by disentangling edits for grammatical error correction. *CoRR*, abs/2407.00934.

Tianyu Yu, Chengyue Jiang, Chao Lou, Shen Huang, Xiaobin Wang, Wei Liu, Jiong Cai, Yangning Li, Yinghui Li, Kewei Tu, Hai-Tao Zheng, Ningyu Zhang, Pengjun Xie, Fei Huang, and Yong Jiang. 2024. Seqgpt: An out-of-the-box large language model for open domain sequence understanding. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19458–19467. AAAI Press.

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

Ding Zhang, Yinghui Li, Qingyu Zhou, Shirong Ma, Yangning Li, Yunbo Cao, and Hai-Tao Zheng. 2023a. Contextual similarity is more valuable than character similarity: An empirical study for chinese spell checking. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. 2023b. 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 Related Work

In the era of LLMs, considering the superior performance of LLMs (Liu et al., 2022; Dong et al., 2023; Liu et al., 2023; Li et al., 2023a; Huang et al., 2023b; Li et al., 2023c; Yu et al., 2024; Li et al., 2024; Du et al., 2024; Xu et al., 2025; Li et al., 2025a,b,c), researchers have invested lots of energy in studying LLMs for GEC tasks (Li et al., 2022b,a; Zhang et al., 2023a; Ye et al., 2022; Ma et al., 2023; Ye et al., 2024b,a, 2025, 2023b; Li et al., 2025d).

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; Ye et al., 2023a; Huang et al., 2023a; 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., 2023b; 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 (Kaneko and Okazaki, 2023b) propose to improve the GEC performance by letting LLMs predict edit spans. Östling et al. (Östling et al., 2023) and Sottana et al. (Sottana et al., 2023) explore the potential of using LLMs as evaluators for English and Swedish GEC tasks. Song et al. (Song et al., 2023) and Kaneko and Okazaki (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.

B Implementation Details and Hyperparameters

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 D.1. The hyperparameter values of the small models to be enhanced in our experiments are shown in Table 4. Besides, the loss functions for Seq2Seq models are the label-smoothed cross-entropy, and the loss function for Seq2Edit is cross-entropy.

C Prompt of LLMs as Corrector

To enable the LLM to directly provide corrected versions of the original sentences, we used the following prompt:

请你针对给出的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合理且忠实的纠正。

例如：

SOURCE SENTENCE 纠正为：TARGET SENTENCE

请你纠正（直接输出纠正后的句子，无需任何解释）：

SOURCE SENTENCE

D Results and Analysis of LLMs as Corrector

Results In Table 1, we observe that in the zero-shot scenario, GPT-3.5-Turbo scores 24.69 and 23.71 for Word-Level and Character-Level $F_{0.5}$, respectively, while `Qwen-72B-Chat` scores 28.15 and 26.51. Under our proposed SEE evaluation method, the $F_{0.5}$ scores for `GPT-3.5-Turbo` and `Qwen-72B-Chat` are 46.51 and 56.76, respectively. Figure 4 and Figure 4 show that in the few-shot scenario, both `GPT-3.5-Turbo` and `Qwen-72B-Chat` improve their scores at the Word-Level and Character-Level.

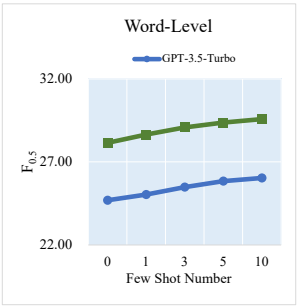

Figure 4: Few-shot results of LLMs on the word-level metric.

Analysis From these experimental results, it is evident that even with the enhancement provided by few-shot learning, there remains a significant gap in the correction capabilities of LLMs. Despite their strong language generation abilities, current LLMs score lower than smaller models under traditional evaluation metrics, which does not align with human perception, as seen in Figure 3. However, our SEE method maintains a high level of alignment with human judgment.

D.1 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 6, and the prompts for evaluation are as shown in Figure 7.

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.

D.2 Case Observation

To verify the correctness of our motivation for using LLMs as explainers, and to demonstrate the explanation information generated by EXAM,

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

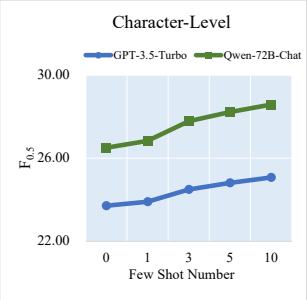

Figure 5: Few-shot results of LLMs on the character-level metric.

we give cases in Table 5 of `GPT-3.5-Turbo` and `Qwen-72B-Chat` acting as the explainer respectively. We can see from Table 5 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.

D.3 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., 2022) has 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.

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

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 6: 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 7: Our designed evaluation prompt of SEE.