ScholarGEC: Enhancing Controllability of Large Language Model for Chinese Academic Grammatical Error Correction

Zixiao Kong[1,2], **Xianquan Wang**[1,2], **Shuanghong Shen**[2*], **Keyu Zhu**[1,2], **Huibo Xu**[1], **Yu Su**[2,3]

[1]Institute of Advanced Technology & State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China

[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

[3]School of Computer Science and Artificial Intelligence, Hefei Normal University

{kzix, wxqcn, kyzhu3, xhbxhb}@mail.ustc.edu.cn, shshen@iai.ustc.edu.cn, yusu@hfnu.edu.cn

Abstract

Large language models (LLMs) have demonstrated exceptional error detection capabilities and can correct sentences with high fluency in grammatical error correction (GEC) tasks. However, when correcting Chinese academic papers, LLMs face significant challenges of over-correction. To delve deeper into this issue, we explore the underlying reasons. On one hand, each discipline has its unique vocabulary and expressions, and LLMs have insufficient and incomplete understanding of domain-specific sentences. On the other hand, the controllability of generative LLMs in GEC tasks is inherently poor, and the traditional sequence-to-sequence (Seq2Seq) correction structure exacerbates this issue. Considering the two aforementioned factors, we propose a new error correction framework for Chinese academic GEC tasks using LLMs, named ScholarGEC. To improve LLMs' understanding of domain-specific knowledge, we construct appropriate disciplinary knowledge prefixes for sentences and use this domain-specific knowledge data to fine-tune the LLM. To enhance the controllability of LLMs, we replace the traditional Seq2Seq structure with a Detection-Correction separated structure. We also introduce a special token during the process to improve the model's error detection stability. Additionally, we incorporate iterative self-reflection to enhance the stability of the generation, in the three parts of LLM generation. Extensive experiments demonstrate the effectiveness and robustness of our framework on a Chinese GEC dataset composed of academic papers, and further analysis reveals the capabilities of our framework in enhancing LLM performance in general GEC tasks.

Code — https://github.com/kzi2000/ScholarGEC

Introduction

According to reports, nearly one million master's and doctoral students graduate annually in China, requiring academic achievements and theses to obtain their degrees. Chinese universities generally mandate that these theses be written in Chinese. Given the sheer volume of theses, relying solely on manual error correction is highly inefficient and often results in missed errors, false positives, and subpar correction quality. Applying Chinese automatic error correction technology to the academic thesis review process can enhance correction efficiency and quality, alleviate the workload of teachers and reviewers, and ensure the rigor and reliability of academic achievements.

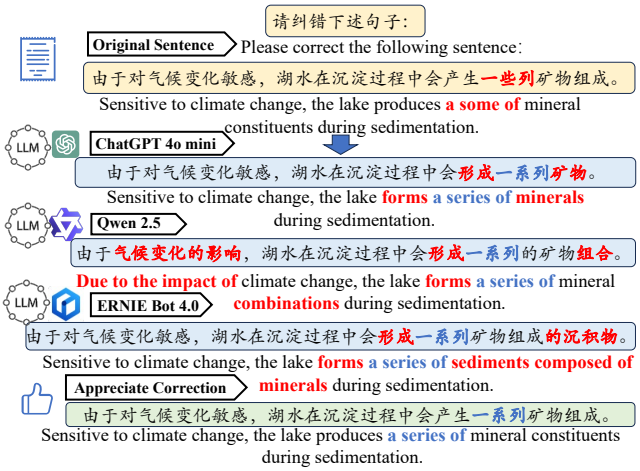

Figure 1: LLMs often make unnecessary or even inappropriate changes in the Chinese Academic grammatical error correction tasks. In the responses of LLMs[1], red text indicates unnecessary or inappropriate changes, and blue text indicates appropriate changes. Also, "mineral constituent" is a technical term in geology and should not be modified.

Chinese Grammatical Error Correction (CGEC) (Zhao et al. 2018) aims to identify and correct potential grammatical errors in a given Chinese sentence while adhering to the principle of minimal edits (Wang et al. 2021). Recently, decoder-only LLMs, with their powerful semantic understanding and logical reasoning capabilities (Zhao et al. 2024; Li et al. 2024), have demonstrated breakthrough performance in various Natural Language Processing (NLP) tasks, showing great potential in CGEC (Fang et al. 2023; Qu and Wu 2023). However, generative LLMs suffer from

*Corresponding author

[1]Inevitably, LLMs' responses are not always exactly the same. However, over-correction occurs frequently.

severe over-correction problems, leading to the unnecessary modification of error-free characters in the source sentence (Wu et al. 2023). As illustrated in Figure 1, the "mineral constituent" is a geological term, but LLMs mistakenly modify it and its collocations. LLMs tend to replace texts with common expression, resulting in unnecessary or incorrect corrections (Park et al. 2020).

Further exploring the phenomenon of over-correction in the CGEC task of academic papers, there are two main reasons for its occurrence: On one hand, there is an insufficient understanding of the semantic information in the text to be corrected, especially evident in sentences that involve strong domain-specific knowledge. Academic papers concentrate on specific issues within particular disciplines or interdisciplinary areas, each with its unique vocabulary and modes of expression. These aspects are essential for professional communication but add complexity to the task of error correction using LLMs. On the other hand, there is the inherent poor controllability of generative LLMs in text correction tasks. Generative LLMs typically adopt the traditional Seq2Seq approach used in GEC tasks. This approach requires the model to generate a mixture of unchanged and corrected parts of the source sentence, which exacerbates the shortcomings associated with poor controllability.

In order to enhance LLMs' adaptability for Chinese academic GEC task, we proposed a framework named ScholarGEC. Integrating additional auxiliary information into tasks is an effective way to improve model's capabilities (Zhang et al. 2022a; Liu et al. 2019; Wang et al. 2023). To enable LLMs to integrate disciplinary knowledge into GEC tasks, we trained a prefix generator to add relevant disciplinary knowledge to source sentences before correction. To increase LLMs' controllability, we introduced a special token. Model avoids unnecessary repetitions by outputting this special token, when a sentence is entirely correct. We also incorporated iterative self-reflection (Piché et al. 2024; Ji et al. 2023) into three parts which utilize generative LLMs, to improve generation quality. We created a dataset called AcaCGEC, using master's and doctoral theses to train the LLM. Experiments on the CGEC benchmark show improved correction abilities and better controllability, reducing over-correction by the LLM. In summary, our main contributions are as follows:

- We proposed a novel approach that integrates disciplinary knowledge into the text correction tasks of LLMs, demonstrating significant effectiveness in the correction of professional academic texts.
- We designed a framework that enhances the error detection capabilities of LLMs, improving the controllability of the detection-correction structure in text correction tasks, and effectively reducing the phenomenon of over-correction.
- We incorporated iterative self-reflection into the text correction tasks of LLMs and demonstrated its effectiveness in improving the accuracy and controllability of LLM text corrections.

Related Work

Traditional Seq2Seq GEC Methods: Traditional CGEC methods often follow the approaches used in English GEC, which are broadly divided into Seq2Edit and Seq2Seq methods (Bryant et al. 2023). Among them, traditional sequence-to-sequence (Seq2Seq) methods (Zhao et al. 2019; Kaneko et al. 2020; Zhang et al. 2022b) can be considered the predecessors of current LLMs for GEC tasks. These methods use encoder-decoder models inspired by neural machine translation to model GEC tasks, where the encoder encodes the source sentence, and the decoder generates target tokens sequentially. Kaneko et al. (2020) further applied pre-trained knowledge to the encoder-decoder model, and Zhang et al. (2022b) explored incorporating grammatical information.

LLMs for GEC: With the success of LLMs in various NLP tasks, researchers have explored their potential in CGEC. Recent studies (Fang et al. 2023; Li et al. 2023; Qu and Wu 2023) evaluated various LLMs (including closed-source and open-source models) on CGEC tasks. Fang et al. (2023) assessed ChatGPT's performance on CGEC through in-context learning, highlighting its ability to generate fluent sentences and its sensitivity to over-correction. Fan et al. (2023) explore open-source LLMs for CGEC via instruction tuning (Ouyang et al. 2022). Zhang et al. (2023) showed that fine-tuned LLMs still struggle to match the performance of existing state-of-the-art lightweight GEC models. These works indicate that the limited success of LLMs in GEC tasks is mainly due to the continuation of traditional Seq2Seq patterns. Howerve, these studies often overlook the issue of over-correction.

Controllability of LLMs for GEC: Seq2Seq models tend to generate sentences with higher probabilities, replacing less common words with more frequent ones, leading to over-correction. Recent studies (Li et al. 2023; Li and Wang 2024; Yang and Quan 2024) have explored methods to alleviate this problem. Li et al. (2023) proposed a two-stage approach, integrating the detection results of Seq2Edit models into Seq2Seq correction models. Due to architectural differences, Li and Wang (2024) integrated detection and correction into a single model and designed a multi-task training method for this structure. Yang and Quan (2024) employed an alignment-enhanced correction approach to alleviate the over-correction problem in LLMs. While these methods mitigate over-correction to some extent, they have several drawbacks: 1) They are often significantly limited in performing GEC tasks on domain-specific corpora. 2) They primarily remediate over-correction problems after they occur, meaning these methods do not fundamentally address the poor controllability of LLMs. Our method differs from existing solutions by effectively integrating domain-specific knowledge into LLM GEC tasks while focusing on fundamentally controlling the occurrence of over-correction.

Method

We propose a framework named ScholarGEC, to enhance the capabilities of LLMs in Chinese academic texts GEC task, with the overall workflow of our approach illustrated in Figure 2. **Firstly**, we design a discipline knowledge in-

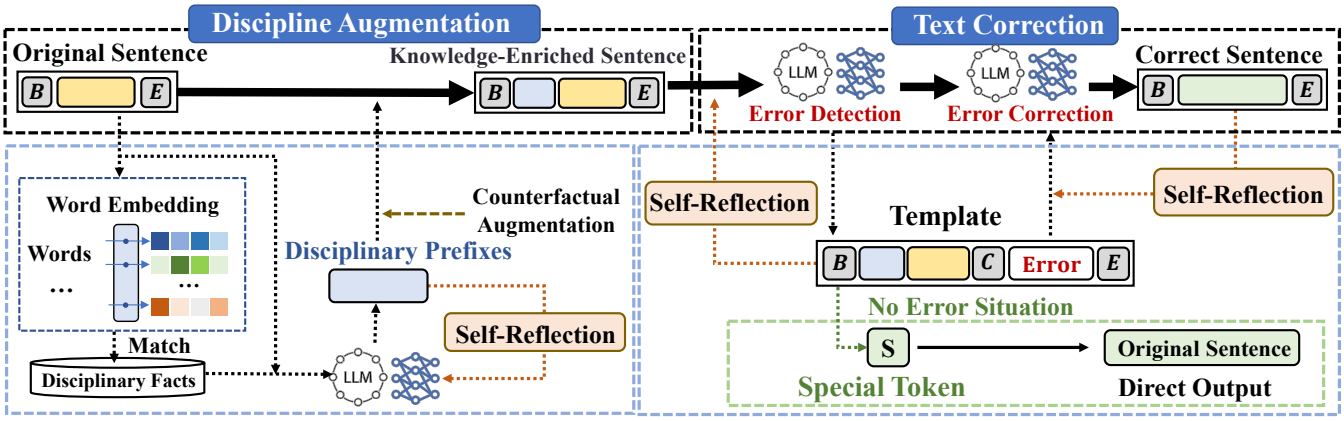

Figure 2: An overview of our framework, which comprises three main components. First, during the **Knowledge Integration** stage, we construct Disciplinary Prefixes that contain relevant knowledge based on the source sentences. Second, in the Detection and Correction phase, we utilize fine-tuned LLMs for Error **Detection and Correction** separately, and we train a Special Token for source sentences that are free of errors, as illustrated in the green section of the figure. Third, we have added **Self-Reflection** mechanisms at three points within the generative LLM, as indicated by the blue sections.

tegration method that enables LLM to perform adaptive error detection and correction using a structured domain-specific knowledge base. **Secondly**, we divide the GEC process into two stages: Error Detection and Error Correction. In the Detection stage, we introduce special tokens to enhance the LLM's ability to identify sentences without errors, thereby skipping the redundant and unpredictable generative correction process. **Finally**, we introduce an iterative Self-Reflection mechanism that involves generating, providing feedback, and refining in a cyclic manner. Through these methods, we aim to improve the performance of LLMs in handling specialized academic texts.

Discipline Knowledge Integration We designe a text Discipline Augmentation method to enable LLM to perform domain-adaptive GEC for Chinese academic papers. The workflow is shown in the Knowledge Integration section of Figure 2. We have integrated a structured domain knowledge base, which contains multiple academic disciplinary facts. Before executing the GEC task, we map both the source sentence and the disciplinary facts from the knowledge base into the same semantic space, followed by calculating the cosine similarity, to match the source sentence with the most semantically similar disciplinary fact. This retrieved fact serves as supplementary material for correcting the source sentence.

We use the LLM to perform word embeddings, mapping both the source sentence and the knowledge base facts into the same vector space. We can define the LLM mapping function f as follows:

$$f : \text{Text} \rightarrow \mathbb{R}^d, \tag{1}$$

where $\mathbb{R}$ is the vector space, d is the dimension of the vector space. For the processing of the source sentence x_s and the domain knowledge y_s from the knowledge base, we can represent them as follows:

$$X_s = f_{\text{LLM}}(x_s)\texttt{[CLS]}, \tag{2}$$

$$Y_s = f_{\text{LLM}}(y_s)\texttt{[CLS]}, \tag{3}$$

where X_s and Y_s are the representations of the source sentence and the domain knowledge in the vector space, respectively. We calculate the cosine similarity between the source sentence and each domain knowledge fact in the knowledge base as Equation below:

$$\text{cosine similarity} = \frac{X_s \cdot Y_s}{\|X_s\| \cdot \|Y_s\|}. \tag{4}$$

However, even with a sufficiently large knowledge base, it is challenging to find enough supporting facts to analyze every sentence in an academic paper (Ni et al. 2024). Therefore, to transform the initially retrieved domain knowledge sentences into more useful domain prefixes, we trained a generator, which is an LLM connected to an external network knowledge base. It takes both the collected domain facts and the source sentence as input to generate appropriate domain prefixes $[p_s]$ through a Knowledge-Aware Reasoning approach (Wu et al. 2024). These domain prefixes, together with the source sentence, form the complete input content for the error detection model.

Domain prefix $[p_s]$ will be concatenated with x_s to form the complete input for the correction model. The actual input tokens are in the form of:

$$x'_s = [p_s^1 p_s^2 ... p_s^m] x_s^1 x_s^2 ... x_s^n, \tag{5}$$

where p_s^i are the tokens of the domain prefix $[p_s]$. where x_s^i are the tokens of the source sentence x_s.

Counterfactual generation aims to eliminate spurious correlations in data. Recent studies have attempted to use counterfactual generation to enhance the robustness of models (Temraz and Keane 2022). During the construction of domain prefixes, we utilized counterfactual augmentation, as shown in Figure 2. We introduced intentional errors into some domain prefixes generated by the LLM. Specifically, within the prefix generator's iterative self-reflection process

(which will be detailed later), terminating the loop with a certain probability ρ and introducing text errors at that point. This approach significantly increased the training data's volume and indirectly trained the LLM to learn from domain-specific data.

Enhance LLMs' controllability The GEC process can be divided into two stages: detection and correction (Qu and Wu 2023; Coyne et al. 2023). In the detection stage, LLM identifies discrepancies in the sentence and outputs the detected information. Based on this information, a template will be constructed, where the erroneous parts are masked for correction. In the correction stage, given a sentence with MASK, the LLM uses autoregressive blank filling (Du et al. 2021) to generate the appropriate segments for each MASK position. In Figure 3, a comparison is made between traditional GEC methods and the Detection-Correction structure.

Given the tokens of the source sentence as follows:

$$x_s = x_s^1 x_s^2 ... x_s^n, \tag{6}$$

the goal of error detection is to predict the detection labels derived from the alignment between the source text and the target text:

$$d = d_1 d_2 ... d_n, d_i \in L = \{R, E\}, \tag{7}$$

where d_i is the detection label for each token, R is the label indicating that the token is correct and requires no modification, and E is the label indicating that the token should be modified.

The training objective for error detection is given by Equation below:

$$\mathcal{L}_D = -\alpha_D (1 - p_\theta(d \mid x_s))^\gamma \log(p_\theta(d \mid x_s)), \tag{8}$$

where θ represents the model parameters, and γ is a hyperparameter set to 2. α_D denotes the weight factor corresponding to the detection labels.

For sentences that do not require modification, we attempt to introduce a special token S to indicate. We hope that LLM can output this special token, during the error detection stage when encountering correct text, thus skipping the error correction process. This is illustrated in Figure 2, where the "No Error Situation" corresponds to the sentences detected as error-free. These sentences will directly output a special token, indicating the model should output the original sentence without any modification. We introduce a new label set $L' = \{R, E, S\}$, where S indicates the sentence is error-free. We define the updated loss function as Equation below:

$$\begin{aligned} \mathcal{L}'_D = & -\alpha_S (1 - p_\theta(S \mid x_s))^\gamma \log(p_\theta(S \mid x_s)) \\ & - \sum_{i=1}^{n} \alpha_D (1 - p_\theta(d_i \mid x_s^i))^\gamma \log(p_\theta(d_i \mid x_s^i)), \end{aligned} \tag{9}$$

where α_S is the weight factor for the special token S. Where $p_\theta(S \mid x_s)$ represents the probability of predicting S given the source sentence x_s. The second term in Equation 9 extends the objective of Equation 8 to handle each token in the sentence. Here, $p_\theta(d_i \mid x_s^i)$ is the probability of predicting the detection label d_i for each token x_s^i. We set a higher weight factor α_S than α_D, to make the model more likely to predict

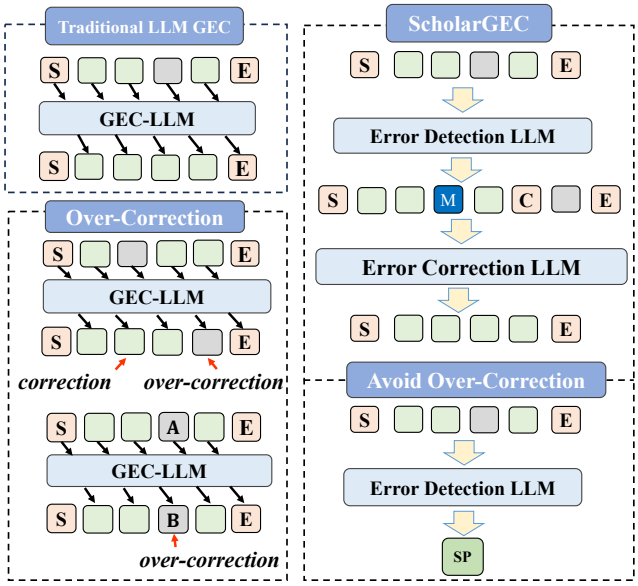

Figure 3: The right side shows the Detection-Correction structure, which has stronger generation controllability compared to the traditional Seq2Seq structure on the left. Additionally, we enhance the error detection model's sensitivity to error-free sentences by training it with a Special Token, highlighted in green in the figure.

S when the sentence is error-free. When the model predicts that the entire sentence is error-free, the first term of the loss function will dominate, prompting the model to output the special token S to indicate that the entire sentence is correct, as illustrated in Figure 3, thus skipping the token-by-token detection. Compared to traditional GEC tasks using LLM, which focus more on correcting text errors, our method emphasizes the model's recognition of text correctness during the error detection stage.

Iterative Self-Reflection We utilized the LLM Self-Reflection method (Ji et al. 2023) multiple times during the correction process. It's an iterative self-refinement approach that alternates between two generative steps (FEEDBACK and REFINE) (Piché et al. 2024). These steps work together to produce high-quality outputs. Self-Reflection relies on an appropriate language model and three prompts (initial generation, feedback, and refinement) and does not require training. Next, we describe Self-Reflection in detail.

After obtaining the disciplinary fact knowledge y_s from the disciplinary knowledge base, the prefix generator initially generates the appropriate disciplinary prefix $[p_s^0]$ as Equation below:

$$[p_s]^0 = \text{LLM}_{\text{generator}}(y_s, x_s), \tag{10}$$

where the superscript zero on $[p_s]^0$ denotes that this domain prefix $[p_s]$ is generated initially. Following the generation of the model's output, it is passed back to the generator for feedback on the previously generated content. This feedback is then used to refine the initial draft.

Model	Parameter	Framework	AcaCGEC-*test*			ECSpell-*test*			NaCGEC-*test*		
			P↑	R↑	$F_{0.5}$↑	P↑	R↑	$F_{0.5}$↑	P↑	R↑	$F_{0.5}$↑
BART	**400M**	**-**	9.53	22.01	10.75	24.85	32.69	26.10	34.67	41.88	35.91
Llama3	**8B**	**-**	18.93	38.19	21.05	35.75	74.25	39.89	37.22	**68.16**	40.94
		ScholarGEC	**29.38**	**46.71**	**31.73**	**41.14**	**77.90**	**45.52**	**45.19**	64.66	**48.09**
Baichuan2	**7B**	**-**	19.76	36.94	21.79	26.54	**80.07**	30.63	33.59	60.49	36.87
		ScholarGEC	**28.58**	**42.31**	**30.56**	**29.10**	74.98	**33.16**	**37.86**	**62.92**	**41.14**
Qwen2	**7B**	**-**	22.74	47.60	25.39	32.77	**71.54**	36.75	35.16	**55.94**	39.31
		ScholarGEC	**33.48**	**52.12**	**36.06**	**37.56**	70.28	**41.42**	**40.26**	54.73	**42.51**
Qwen2.5	**14B**	**-**	25.59	49.63	28.34	35.62	73.05	39.69	38.59	58.28	41.39
		ScholarGEC	**37.81**	**55.26**	**40.36**	**42.09**	**73.92**	**46.06**	**45.08**	**59.33**	**47.35**

Table 1: Results on AcaCGEC-*test*, ECSpell-*test* and NaCGEC-*test*. BART (Lewis et al. 2019) is a traditional Seq2Seq model, while Llama3, Baichuan2, Qwen2 and Qwen2.5 are LLMs pre-trained for Chinese. In the Framework column, the dash **-** indicates no framework is used. The data for ECSpell-*test* and NaCGEC-*test* are subsets of the full test data after cleaning and proportion adjustment. Bolded results represent the better performer between two rows for each individual model.

Similarly, during both the error detection and correction processes, the results generated by the model are fed back into LLM for feedback and refinement. In this iterative self-reflection process, the model evaluates and provides feedback on the previously generated content, based on the feedback prompt, and subsequently refines the draft, repeating this process. We employ few-shot prompting (Brown et al. 2020) to guide the model in generating feedback and incorporating it into the revised draft. The self-reflective process for acquiring domain facts is described by the following equations:

$$[feed]^i = \text{LLM}_{\text{feedback}}([p_s]^i), \tag{11}$$

$$[p_s]^{i+1} = \text{LLM}_{\text{refine}}([p_s]^i, [feed]^i), \tag{12}$$

where $[feed]^i$ is the feedback for the i round of generation. Subsequently, the feedback $[feed]^i$ will be fed back to refine the previous generation result $[p_s]^i$, with the refined result $[p_s]^{i+1}$ becoming the new generation result.

Experiment

Experimental Setup

Dataset and Evaluation Metrics We constructed a dataset called AcaCGEC, which is derived from the theses of approximately 1,000 master's and doctoral students. We manually extracted sentences closely related to domain-specific knowledge from the data. After expert annotation, we obtained 51,412 sentence pairs containing academic domain knowledge. Additionally, as mentioned before, we used counterfactual augmentation during the training phase, which increased the amount of usable training data.

To evaluate the model's GEC capabilities for Chinese academic papers, we assess the model on AcaCGEC-*test*, NaCGEC-*test* (Ma et al. 2022), and ECSpell-*test* (Lv et al. 2023). We report precision, recall and $F_{0.5}$ score for all experiments, and the $F_{0.5}$ score has been found to have a better correlation with human judgment compared to other metrics (Grundkiewicz, Junczys-Dowmunt, and Gillian 2015; Napoles et al. 2015; Chollampatt and Ng 2018), it can be considered that this parameter is most suitable to be used to evaluate the GEC capability of the model.

Constructing the Disciplinary Knowledge Base Disciplinary knowledge sentences serve as the initial input for the prefix generator. Since most of the content in our AcaCGEC dataset comes from natural sciences, mathematics, engineering, and humanities and social sciences, We obtained most of the key concept definitions from the mainstream textbooks used in these disciplines in higher education institutions in mainland China, totaling 1,000 sentences, as the disciplinary knowledge base.

Implementation Process To verify the effectiveness of our framework, we chose the Llama3-8B-Chinese-Chat model (Wang et al. 2024), Baichuan2-7B-Chat (Yang et al. 2023), Qwen2-7B-Instruct and Qwen2.5-14B-Instruct (Yang et al. 2024) as the base models. The Llama3-8B-Chinese-Chat model is built on the Meta-Llama-3-8B-Instruct model (Dubey et al. 2024). For the error detection and correction part, we applied LoRA fine-tuning (Hu et al. 2021) to the LLMs using the training dataset, with a LoRA scaling factor alpha set to 16, a rank set to 8, and fine-tuning applied to all layers. We performed fine-tuning training on the complete training set for 10 epochs, separately for the model's error detection and correction capabilities.

Performance Comparisons

Main Result By comparing the metric results in adjacent rows of Table 1, we find that compared to LLMs without using any framework, ScholarGEC significantly improves the $F_{0.5}$ score across all datasets. Specifically, the improvement on AcaCGEC-*test* dataset is approximately 10 percentage points higher than on other datasets. This demonstrates the effectiveness of our framework in enhancing the GEC capabilities of LLMs, particularly for corpora containing academic knowledge. A notable feature of the experimental results is that the recall is much higher than the precision. Upon analyzing specific samples, we find that this is because the models perform many correct corrections but also make

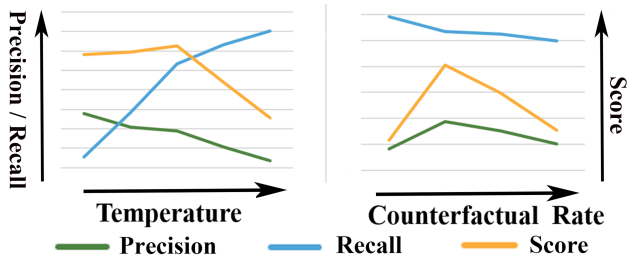

Figure 4: Hyperparameter analysis on Llama3-8B-Chinese-Chat with the ScholarGEC framework. We examined the temperature τ of LLM generation and the counterfactual sample ratio ρ employed in our counterfactual augmentation process. **Score** in the figure means $F_{0.5}$ score.

numerous over-corrections, which lowers the precision. After applying ScholarGEC to LLMs, there is a particularly significant improvement in precision, while the increase in recall is not as noticeable. This suggests that our framework enhances the GEC capability of the models while mitigating over-correction to some extent. However, even without the use of a framework, the latest LLMs outperform the traditional Seq2Seq structure of the BART (Lewis et al. 2019) model in CGEC tasks, especially in datasets heavily related to professional knowledge. This can be attributed to the extensive pretraining of LLMs on large datasets and possibly due to structural differences; traditional CGEC methods are adapted from English GEC, which may result in poorer performance when dealing with complex Chinese text. In addition, models with larger parameters generally perform better.

In addition, we observe several interesting facts. First, the performance improvement on the AcaCGEC-*test* dataset is significantly more pronounced than on other datasets. This can be attributed to the fact that the NaCGEC-*test* and ECSpell-*test* datasets mainly consist of materials that generally lack professional content. As a result, our knowledge integration mechanism does not play a significant role, and we will validate this conclusion through subsequent ablation experiments. Second, in some experiments, the Recall metric decreased rather than increased after the LLMs utilized the framework. This may be due to the Detection-Correction structure and the iterative self-reflection we adopted, which make the LLMs more conservative in making modifications. Third, the Qwen model generally performed well on AcaCGEC, but on other datasets, the $F_{0.5}$ score obtained by the Qwen2.5-14B is sometimes even worse than that of Llama3-8B. We speculate that this is due to the differences in the pretraining corpora of the Qwen series models compared to Llama3, giving it an advantage when dealing with highly specialized texts.

Hyperparameter Study We analyzed the impact of hyperparameter selection on the Llama3-8B-Chinese-Chat model. The results are shown in Figure 4. The temperature controls the randomness and diversity of the output. Lower temperatures were found to decrease the Recall. This is likely because professional knowledge often involves many uncommon technical terms and a lower temperature favors higher-probability words. The counterfactual ratio ρ determines the proportion of samples undergoing counterfactual augmentation. As shown in the right part of Figure 4, a lower ratio yields higher Recall but much lower Precision. This also indicates the effectiveness of counterfactual augmentation in mitigating over-correction.

Dataset	Average Length	Method	Time↓
ECSpell	41.77	-	34.54
		ScholarGEC	**25.62**
AcaCGEC	37.94	-	51.43
		ScholarGEC	32.11
		w/o Self-Reflection	**21.92**
		w/o Special Token	64.95

Table 2: Efficiency Analysis on the Llama3-8B-Chinese-Chat, using two NVIDIA RTX 4090 GPUs. We selected 50 sentences as a group for each dataset. **Average Length** represents the average sentence length per group, measured in tokens. **Time** represents the overall time taken for processing each group, measured in seconds. w/o means "without".

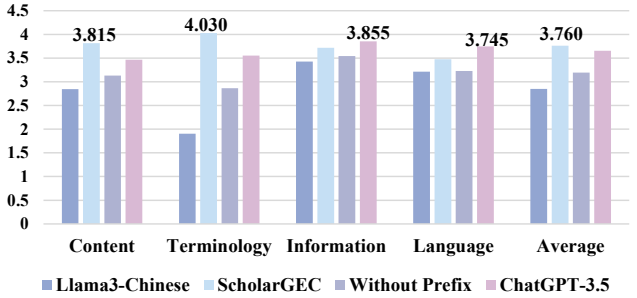

Figure 5: Human Evaluation Study focusing on four aspects of the corrected text: **Content** Consistency, **Terminology** Consistency, **Information** Completeness, and **Language** Accuracy. Participants were selected to rate these aspects on a scale from 1 to 5, where each metric represents the average score obtained from all. **ScholarGEC** in figure is based on Llama3-Chinese, and the **Without Prefix** means ScholarGEC without knowledge prefix construction. Bolded results represent the SOTA.

Efficiency Analysis In practical applications of LLMs for the GEC task, it is important to consider not only accuracy but also inference time to meet real-time user needs. We examined the impact of our framework on LLM inference times, as shown in Table 2. The framework generally reduces the time required for inference. Further ablation studies revealed that, although the Self-Reflection mechanism increases processing time, the use of Special Tokens significantly cuts down overall time, by avoiding redundant generations of error-free sentences.

Human Evaluation Study Given the ambiguity in Chinese, it is difficult to find a suitable quantitative metric to evaluate the extent of over-correction in generated results. Therefore, we validated the model's performance of handling over-correction through human evaluation studies. Fol-

Pre-trained	KI	DC	SP	SR	AcaCGEC-*test* P↑	AcaCGEC-*test* R↑	AcaCGEC-*test* $F_{0.5}$↑	ECSpell-*test* P↑	ECSpell-*test* R↑	ECSpell-*test* $F_{0.5}$↑
Yes	✓	✓	✓	✓	<u>29.38</u>	**46.71**	**31.73**	**41.14**	<u>77.90</u>	**45.52**
Yes	✓	✓	✓	×	27.08	<u>46.16</u>	27.60	38.74	77.64	43.05
Yes	✓	✓	×	✓	**29.76**	42.20	<u>31.62</u>	40.45	75.01	44.56
Yes	×	✓	✓	✓	25.64	43.81	27.96	<u>40.92</u>	**78.32**	<u>45.24</u>
Yes	×	✓	✓	×	24.49	46.01	27.02	39.25	77.09	43.52
Yes	✓	×	×	✓	26.73	42.77	28.90	36.69	74.31	40.82
Yes	×	×	×	×	23.93	41.73	26.16	36.22	74.97	40.39
No	✓	✓	✓	✓	11.07	40.13	12.94	37.79	76.62	42.05
No	×	×	×	×	18.93	38.19	21.05	35.75	74.25	39.89

Table 3: Ablation experiments on the ScholarGEC framework using the Llama3-8B-Chinese-Chat as the base model. **KI** stands for Knowledge Integration, **DC** indicates the use of Detection-Correction architecture, **SP** represents Special Token training, and **SR** denotes the LLM's iterative Self-Reflection mechanism. The best results are highlighted in bold, while the second-best results are underlined.

lowing key concepts in Chinese grammar research, we used the following metrics:

- **Content Consistency**: How does the consistency of core ideas and conclusions in the modified text?
- **Terminology Consistency**: How does the consistency of professional terminology in the modified text?
- **Information Completeness**: How does the completeness of information and data in the modified text?
- **Language Accuracy**: How does the accuracy of language in the modified text?

We invited twenty volunteers to evaluate the correction results of various methods on AcaCGEC. The results are shown in Figure 5, the LLM using ScholarGEC shows improvements in the average scores across all metrics, with a particularly significant increase in the Terminology. This demonstrates the effectiveness of our framework in enhancing the LLMs' controllability in academic GEC task. Additionally, we removed the knowledge prefix mechanism from the complete framework, as indicated by the w/o KI in the figure. As a result, the score in Terminology is much lower than the complete ScholarGEC, validating the effectiveness of supplementing LLMs with disciplinary knowledge. Meanwhile, the Llama3-Chinese model using ScholarGEC, with far fewer parameters than ChatGPT-3.5, achieved significantly higher scores in Terminology and Content, and set a new SOTA in Average score.

Ablation Study

To investigate the effectiveness of various components within our framework, we conducted extensive experimental evaluations, using different configurations of the framework. The framework configurations and corresponding experimental results are shown in Table 3.

Fine-tuning Effectiveness Models without pretraining showed significantly lower performance on AcaCGEC-*test* compared to those with pretraining, even under identical framework configurations. Furthermore, when the framework was applied directly to an un-finetuned LLM, the performance on AcaCGEC-*test* actually decreased. This suggests that mechanisms specifically designed for domain knowledge, such as constructing Knowledge Prefixes, may introduce instability when the model lacks sufficient understanding of the domain. On the other hand, the metrics on ECSpell-*test* improved, supporting this hypothesis.

Effectiveness of Knowledge Prefix In Table 3, removing the Knowledge Prefix mechanism resulted in a noticeable decrease in the $F_{0.5}$ score on AcaCGEC-*test*, while there was almost no change on ECSpell-*test*. This confirms the unique effectiveness of the Knowledge Prefix mechanism in enhancing the LLM's understanding of domain-specific knowledge.

Effectiveness of Special Tokens When Special Tokens were removed and the Detection-Correction structure was replaced with a Seq2Seq structure, the $F_{0.5}$ score decreased on both datasets. However, the model performed well on ECSpell-*test* with the DC structure, which further validates the positive impact of this architecture on the model's capabilities. Additionally, when Special Tokens were removed, although the model achieved the highest precision on AcaCGEC-*test*, the change was not significant, while the recall dropped significantly, suggesting that Special Tokens can enhance the model's ability to detect errors, at the cost of perhaps slightly reducing the precision of corrections.

Effectiveness of Self-Reflection After removing the Self-Reflection structure, we observed a decrease in performance across all metrics, particularly a notable drop in the Precision and $F_{0.5}$ score on AcaCGEC-*test* when comparing the two preceding rows in Table 3. This demonstrates the critical role of the self-reflective structure in the framework, which helps the model generate more accurate answers.

Conclusion

We introduced a novel and highly controllable framework for leveraging domain knowledge in LLM-based Chinese GEC tasks. This framework specifically addressed two main issues that caused poor controllability in LLMs for traditional correction tasks: insufficient understanding of domain knowledge and instability in Seq2Seq structures. Experimental results showed that our framework effectively improved LLMs' error correction capabilities for domain-specific texts. On our dataset constructed from real academic papers, our framework significantly improved LLM's $F_{0.5}$ score. Ablation studies confirmed the effectiveness of our framework, and human evaluation results indicated that our framework significantly mitigated over-correction.

However, our work still has the following limitations: First, due to computational resource constraints, we did not experiment with LLMs with larger parameter counts, which might have performed differently. Second, for the SelfReflection phase, we did not extensively explore prompt engineering efficiency improvements. Third, the improvement in LLM correction abilities by our framework is less pronounced in general domains, compared to academic ones. The future work aims to find a more universally applicable method to enhance the controllability of LLMs when performing GEC tasks.

Acknowledgements

This research was partially supported by grants from the China Postdoctoral Science Foundation (Grant No. 2024M760725), the Anhui Provincial Natural Science Foundation (Grant No. 2408085QF212), the Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (No. MMC202405), and the Fundamental Research Funds for the Central Universities.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Bryant, C.; Yuan, Z.; Qorib, M. R.; Cao, H.; Ng, H. T.; and Briscoe, T. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3): 643–701.

Chollampatt, S.; and Ng, H. T. 2018. A reassessment of reference-based grammatical error correction metrics. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2730–2741.

Coyne, S.; Sakaguchi, K.; Galvan-Sosa, D.; Zock, M.; and Inui, K. 2023. Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction. *arXiv preprint arXiv:2303.14342*.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fan, Y.; Jiang, F.; Li, P.; and Li, H. 2023. Grammargpt: Exploring open-source llms for native chinese grammatical error correction with supervised fine-tuning. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 69–80. Springer.

Fang, T.; Yang, S.; Lan, K.; Wong, D. F.; Hu, J.; Chao, L. S.; and Zhang, Y. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

Grundkiewicz, R.; Junczys-Dowmunt, M.; and Gillian, E. 2015. Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 461–470.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ji, Z.; Yu, T.; Xu, Y.; Lee, N.; Ishii, E.; and Fung, P. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1827–1843.

Kaneko, M.; Mita, M.; Kiyono, S.; Suzuki, J.; and Inui, K. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. *arXiv preprint arXiv:2005.00987*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, R.; Liu, Q.; He, L.; Zhang, Z.; Zhang, H.; Ye, S.; Lu, J.; and Huang, Z. 2024. Optimizing Code Retrieval: High-Quality and Scalable Dataset Annotation through Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2053–2065.

Li, W.; and Wang, H. 2024. Detection-Correction Structure via General Language Model for Grammatical Error Correction. *arXiv preprint arXiv:2405.17804*.

Li, Y.; Liu, X.; Wang, S.; Gong, P.; Wong, D. F.; Gao, Y.; Huang, H.-Y.; and Zhang, M. 2023. TemplateGEC: Improving grammatical error correction with detection template. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6878–6892.

Liu, Q.; Huang, Z.; Yin, Y.; Chen, E.; Xiong, H.; Su, Y.; and Hu, G. 2019. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1): 100–115.

Lv, Q.; Cao, Z.; Geng, L.; Ai, C.; Yan, X.; and Fu, G. 2023. General and Domain-adaptive Chinese Spelling Check with Error-consistent Pretraining. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(5): 1–18.

Ma, S.; Li, Y.; Sun, R.; Zhou, Q.; Huang, S.; Zhang, D.; Yangning, L.; Liu, R.; Li, Z.; Cao, Y.; et al. 2022. Linguistic Rules-Based Corpus Generation for Native Chinese Grammatical Error Correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Napoles, C.; Sakaguchi, K.; Post, M.; and Tetreault, J. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 588–593.

Ni, P.; Wang, X.; Lv, B.; and Wu, L. 2024. GTR: An explainable Graph Topic-aware Recommender for scholarly document. *Electronic Commerce Research and Applications*, 67: 101439.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Park, C.; Yang, Y.; Lee, C.; and Lim, H. 2020. Comparison of the evaluation metrics for neural grammatical error correction with overcorrection. *IEEE Access*, 8: 106264–106272.

Piché, A.; Milios, A.; Bahdanau, D.; and Pal, C. 2024. LLMs can learn self-restraint through iterative self-reflection. *arXiv preprint arXiv:2405.13022*.

Qu, F.; and Wu, Y. 2023. Evaluating the capability of large-scale language models on Chinese grammatical error correction task. *arXiv preprint arXiv:2307.03972*.

Temraz, M.; and Keane, M. T. 2022. Solving the class imbalance problem using a counterfactual method for data augmentation. *Machine Learning with Applications*, 9: 100375.

Wang, J.; Wu, L.; Zhao, H.; and Jia, N. 2023. Multi-view enhanced zero-shot node classification. *Information Processing & Management*, 60(6): 103479.

Wang, S.; Zheng, Y.; Wang, G.; Song, S.; and Huang, G. 2024. Llama3-8B-Chinese-Chat (Revision 6622a23).

Wang, Y.; Wang, Y.; Dang, K.; Liu, J.; and Liu, Z. 2021. A comprehensive survey of grammatical error correction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5): 1–51.

Wu, H.; Wang, W.; Wan, Y.; Jiao, W.; and Lyu, M. 2023. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. *arXiv preprint arXiv:2303.13648*.

Wu, L.; Li, Z.; Zhao, H.; Huang, Z.; Han, Y.; Jiang, J.; and Chen, E. 2024. Supporting Your Idea Reasonably: A Knowledge-Aware Topic Reasoning Strategy for Citation Recommendation. *IEEE Transactions on Knowledge and Data Engineering*.

Yang, A.; Xiao, B.; Wang, B.; Zhang, B.; Bian, C.; Yin, C.; Lv, C.; Pan, D.; Wang, D.; Yan, D.; et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yang, H.; and Quan, X. 2024. Alirector: Alignment-Enhanced Chinese Grammatical Error Corrector. *arXiv preprint arXiv:2402.04601*.

Zhang, K.; Zhang, K.; Zhang, M.; Zhao, H.; Liu, Q.; Wu, W.; and Chen, E. 2022a. Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis. *arXiv preprint arXiv:2203.16369*.

Zhang, Y.; Cui, L.; Cai, D.; Huang, X.; Fang, T.; and Bi, W. 2023. Multi-Task Instruction Tuning of LLaMa for Specific Scenarios: A Preliminary Study on Writing Assistance. arXiv:2305.13225.

Zhang, Y.; Zhang, B.; Li, Z.; Bao, Z.; Li, C.; and Zhang, M. 2022b. SynGEC: Syntax-enhanced grammatical error correction with a tailored GEC-oriented parser. *arXiv preprint arXiv:2210.12484*.

Zhao, H.; Zheng, S.; Wu, L.; Yu, B.; and Wang, J. 2024. Lane: Logic alignment of non-tuning large language models and online recommendation systems for explainable reason generation. *arXiv preprint arXiv:2407.02833*.

Zhao, W.; Wang, L.; Shen, K.; Jia, R.; and Liu, J. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. *arXiv preprint arXiv:1903.00138*.

Zhao, Y.; Jiang, N.; Sun, W.; and Wan, X. 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 7*, 439–445. Springer.