# EXCGEC: A Benchmark of Edit-wise Explainable Chinese Grammatical Error Correction

Anonymous ACL submission


## Abstract

Existing studies explore the explainability of Grammatical Error Correction (GEC) in a limited scenario, where they ignore the interaction between corrections and explanations. To bridge the gap, this paper introduces the task of EXplainable GEC (**EXGEC**), which focuses on the integral role of both correction and explanation tasks. To facilitate the task, we propose **EXCGEC**, a tailored benchmark for Chinese EXGEC consisting of 8,216 explanation-augmented samples featuring the design of hybrid edit-wise explanations. We benchmark several series of LLMs in multiple settings, covering post-explaining and pre-explaining. To promote the development of the task, we introduce a comprehensive suite of automatic metrics and conduct human evaluation experiments to demonstrate the human consistency of the automatic metrics for free-text explanations.[1]


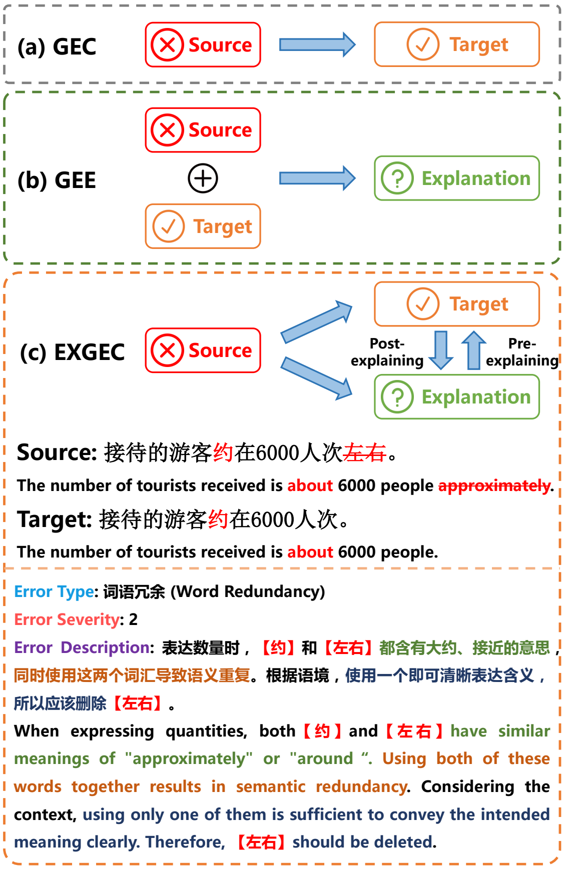


Figure 1: Task definitions of GEC, GEE and EXGEC. For the error description of EXGEC, we highlight **evidence words**, **linguistic knowledge**, **error causes**, and **revision advice** parts in different colors.

## 1 Introduction

Despite the notable advancements in Grammatical Error Correction (GEC) (Bryant et al., 2023; Zhao et al., 2018; Bryant et al., 2019a), there still exists a lack of profound examination into the explainability of GEC (Dwivedi et al., 2023), which is critical in educational scenarios for L2 (Language second)-speakers (Wang et al., 2021) or school-age children (Li et al., 2023b). These mainstream users, who often face challenges in creating grammatically accurate and fluent texts, may be confused or even misguided if they are provided with limited access to only corrective texts. Therefore, augmenting the explainability of GEC is unquestionably beneficial for the progression of the GEC community as well as related fields, such as essay scoring (Ashiya Katuka et al., 2024; Stahl et al., 2024), intelligent tutoring systems (Montenegro-Rueda et al., 2023) and other emerging educational scenarios (Lan et al., 2024).

As illustrated in Figure 1, existing tasks like GEC and Grammatical Error Explanation (GEE) typically address either correction or explanation, ignoring the interaction between the two. To bridge the gap, we introduce the task of **EX**plainable **G**rammatical **E**rror **C**orrection (**EXGEC**). By integrating these two tasks, EXGEC enables systems to elucidate the linguistic knowledge and reasoning mechanism underlying predicted corrections, thereby achieving the best of both worlds. Additionally, EXGEC can function as a test bed for

[1] All the codes and data will be released after the review.

determining the explainable abilities of large language models (LLMs) and identifying any unintended biases and risks in educational scenarios.

To facilitate EXGEC, we present **EXCGEC**, a tailored benchmark for Chinese EXGEC, featuring the design of hybrid edit-wise explanations. Each explanation, based on a particular edit, consists of three elements: 1) *Error types*, which allow learners to absorb syntax and semantic knowledge in an inductive way (Fei et al., 2023). We establish a hierarchical and pragmatic two-tier taxonomy for Chinese grammatical errors. 2) *Error severity levels* ranging from $1 \sim 5$ points, which are beneficial to prioritize core corrections. 3) *Error descriptions*, presented as the form of natural language explanation (Camburu et al., 2018; He et al., 2023), provide evidence words, relevant linguistic knowledge or syntax rules, error causes, and revision advice for edits. The design provides more detailed and faithful guidance for learners, allowing them to comprehend each grammatical error committed. This is unlikely achievable for other designs such as example-based (Kaneko et al., 2022) or sentence-level explanations (Nagata et al., 2021).

Stimulated by recent success of synthetic data generation (Shum et al., 2023; Whitehouse et al., 2023), we employ a semi-automatic dataset construction solution to enhance efficiency, while minimising annotation costs. Initially, we synthesize the EXCGEC dataset by prompting GPT-4 (Liu et al., 2024). Then we hire native annotators to filter invalid data and provide a detailed analysis of invalid data, ensuring the quality of the dataset (Ding et al., 2024). We finally obtain 8,216 clean explanation-augmented samples for benchmarking. We also introduce automatic metrics to evaluate performance across both tasks, and conduct human evaluation experiments to ascertain the correlation between these metrics and human judgement, thus demonstrating their effectiveness.

Based on the benchmark, we develop EXGEC baseline models that can perform both the correction and explanation tasks in either post-explaining (correct-then-explain) or pre-explaining (explain-then-correct) sequences. Particularly, we design **Co**rrect-**T**hen-**E**xplain (**COTE**) decoding algorithm for post-explaining models. Benchmarking various series of open-source LLMs has yielded several intriguing findings. For example, post-explaining models display higher performance than pre-explaining models. Moreover, COTE significantly enhances performance by alleviating the alignment workload for the LLMs. Our contributions in this paper are listed as follows:

(1) We introduce the EXGEC task and establish a corresponding benchmark consisting of a Chinese EXGEC dataset and a comprehensive set of automatic metrics, contributing to the stable development of the field of EXGEC.

(2) We develop EXGEC baseline models and investigate the abilities of various LLMs using our proposed benchmark.

(3) We conduct detailed analyses on our proposed dataset and baselines to gain further insights. Human evaluation experiments are also conducted to confirm the effectiveness of automatic metrics for error descriptions.

## 2 Related Work

Exploration of explainable GEC has witnessed a paradigm shifting from fine-tuning (Kaneko and Okazaki, 2023) to prompting (Zhao et al., 2024), with the focus being local explanations of individual predictions. Fei et al. (2023) construct an explainable GEC dataset EXPECT, which is annotated with evidence words and error types based on the standard GEC benchmark (Bryant et al., 2019b). However, EXPECT falls short of flexibility due to the lack of natural language explanations. To fill the gap, Song et al. (2023) propose the task of grammatical error explanation. They observe that GPT-4 suffers from identifying and explaining errors with limited access to only parallel source-target pairs. To address this issue, they fine-tune an extra LLM as an edit extractor, which is trained on synthesized data. However, all these studies overlook the benefits of effectiveness and efficiency brought by multi-task learning both correction and explanation tasks, which is extensively explored in this work.

On the other hand, a similar task, called feedback comment generation task (Nagata et al., 2021; Hanawa et al., 2021), focuses on sentence-level explanations. However, it suffers from high cost associated with data annotation (Nagata et al., 2020). Furthermore, due to the complexity of the task (Nagata, 2019), it is often explored with limited access to only a subset of grammatical error types.

## 3 Task Definition

### 3.1 Grammatical Error Correction

GEC has been studied for decades, witnessing the shift from rule-based methods to LLM-based methods. Formally, given an ungrammatical text (source text) $X = \{x_1, x_2, \cdots, x_T\}$, a GEC model is required to correct $X$ into a grammatically correct counterpart (target text) $Y = \{y_1, y_2, \cdots, y_{T'}\}$ without changing the original semantic as far as possible. Typically, GEC is usually treated as a sequence-to-sequence (Seq2Seq) task, the training objective of which is formulated as follow:

$$\mathcal{L}_{\text{GEC}} = -\sum_{t=1}^{T'} \log P(y_t \mid Y_{<t}, X) \quad (1)$$

### 3.2 Grammatical Error Explanation

GEE has been explored in several methodologies, including sentence-level explanation and edit-wise explanation. Since sentence-level explanation suffer from over-generalization and confusion especially when a sentence contains multiple grammatical errors, this work focuses solely on edit-wise explanations. Given a source text $X$ and its target counterpart $Y$, the GEE model needs to explain each grammatical error $e_i$ in $X$. Specifically, GEE is typically solved in a two-step pipeline consisting of edit extraction and edit-wise explanation. 1) **Edit extraction** produces an edit set $E = \{e_1, e_2, \cdots, e_n\}$ that represent grammatical errors in $X$ and also clarify the transformation from ungrammatical segments of $X$ to target segments of $Y$. Typically, an edit contains four key elements: source position $sp$, source content $sc$, target position $tp$, and target content $tc$. The process of edit extraction can be easily accomplished using alignment-based evaluation toolkits like ERRANT (Bryant et al., 2017; Felice et al., 2016) and CLEME (Ye et al., 2023). 2) **Edit-wise explanation** generates a set of explanations $E' = \{e'_1, e'_2, \cdots, e'_n\}$, with each explanation $e'_i$ corresponding to $e_i$, given the source and the target texts. Although the design of explanation varies across related work (Song et al., 2023; Zhao et al., 2024), the typical training objective of GEE models is presented as follows:

$$E = f(X, Y) \quad (2)$$

$$\mathcal{L}_{\text{GEE}} = -\sum_{i=1}^{n} \log P(e'_i \mid X, Y, e_i) \quad (3)$$

where $f : (X, Y) \rightarrow E = \{(sp_i, sc_i, tp_i, tc_i)\}_{i=1}^{n}$ is the edit extraction function used to extract edits of $X$ and $Y$, and $n$ is the number of edits.

Existing studies (Song et al., 2023; Fei et al., 2023) focus on developing GEE models that can generate more reasonable explanations. However, an extra GEC model is compulsory to allow GEE models to generate explanations if only source texts are offered, thus resulting in an issue of low efficiency. Furthermore, there exists a gap between GEC and GEE models if they are trained on different data with domain shift.

### 3.3 Explainable Grammatical Error Correction

To get rid of the drawbacks brought by the natures of GEE, we propose the EXGEC task which aims to perform both correction and explanation tasks simultaneously. The motivation of combining these two tasks majorly falls on two aspects. First, a branch of existing studies (Wiegreffe and Marasović, 2021; Hartmann and Sonntag, 2022; Li et al., 2022, 2024) have demonstrated training with access to human explanations can improve model performance. It is also intuitive that either of GEC and GEE tasks can mutually benefit from each other when training in a multi-task manner. Second, it is more time-saving and cost-efficient to deploy a single EXGEC model rather than two detached models in foreign language education platforms.

In this task, the only input element is an ungrammatical source text $X$, and the EXGEC model learns to output both the grammatically target text $Y$ and explanations $E'$. Similar to GEE, EXGEC follows the edit-wise style of explanation, and it is categorized into two different settings by the order of correction and explanation tasks, with the basic scheme of multi-task learning.

**Post-explaining.** Models are trained first to generate target texts (Camburu et al., 2018), which allows the explanations to be explicitly conditioned on the target texts, thus ensuring high faithfulness of explanations towards the target texts. The training objective is as follows:

$$\begin{aligned}\mathcal{L}_{\text{post}} = &-\sum_{t=1}^{T'} \log P(y_t \mid Y_{<t}, X) \\ &-\sum_{i=1}^{n} \log P(e'_i \mid X, Y, e_i)\end{aligned} \quad (4)$$

The inference of post-explaining models is represented as follows:

$$\hat{Y} = \text{EXGEC}_{\text{post}}(X) \quad (5)$$

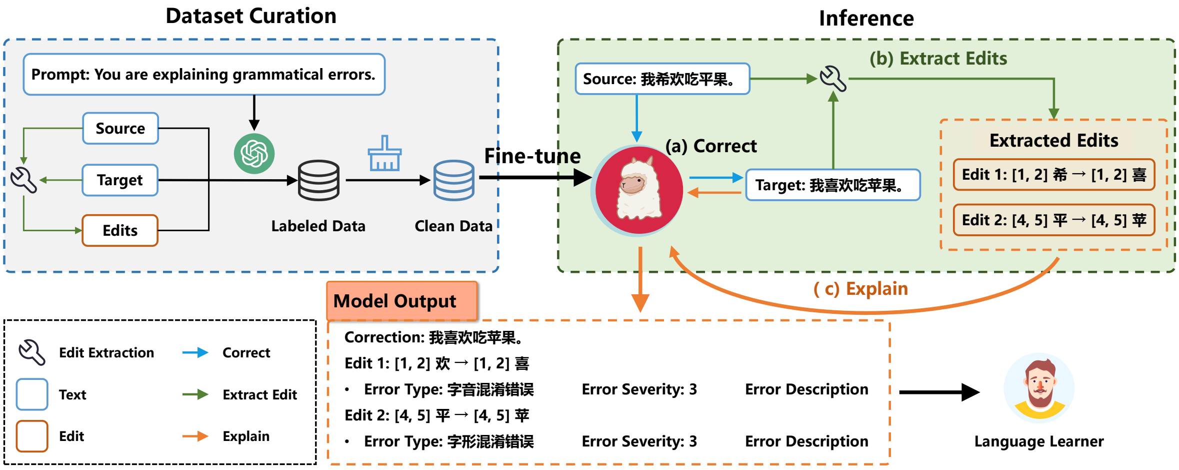


Figure 2: Overview of benchmark construction and model development. We show the inference process of the post-explaining model in particular.

$$\hat{E}' = \text{EXGEC}_{\text{post}}(X, Y, f(X, \hat{Y})) \tag{6}$$

With the target texts generated ahead, post-explaining models can output explanations conditioned on the specific edits that are extracted by aligning the source and the target texts, thus improving accuracy and faithfulness of explanations.

**Pre-explaining.** This type of models are trained conversely, whose mechanism is similar to the Chain of Thought (CoT) technique. Pre-explaining models are supposed to make full use of synthesized explanations to generate elaborated target texts. With minimal modification from Equation (4), the training objective of pre-explaining models is as follow:

$$\begin{aligned}\mathcal{L}_{\text{pre}} = &-\sum_{i=1}^{n} \log P(e'_i \mid X) \\ &-\sum_{t=1}^{T'} \log P(y_t \mid Y_{<t}, X, E')\end{aligned} \tag{7}$$

Notably, pre-explaining models may struggle to generate well-formed edit-wise explanations due to the inaccessibility to the edit extraction function $f$, which necessitate both the source and the target texts. Similarly, the inference of pre-explaining models is presented as follows:

$$\hat{E}' = \text{EXGEC}_{\text{pre}}(X) \tag{8}$$

$$\hat{Y} = \text{EXGEC}_{\text{pre}}(X, E') \tag{9}$$

## 4 EXCGEC Benchmark

To facilitate the development of EXGEC task, we construct EXCGEC, the first benchmark for explainable Chinese GEC particularly. As illustrated in Figure 2, we begin by the process of data curation, which consists of explanation design in Section 4.1, explanation synthesis and refinement in Section 4.2. Then we gain in-depth understanding of GPT-4 (Achiam et al., 2023) in EXGEC by further analyzing generated data in Section 4.3, where we summarize common failure modes in invalid instances. Finally, we introduce a series of automatic metrics for evaluating explanations in Section 4.4.

### 4.1 Explanation Design

In the pursuit of comprehensiveness and plausibility, we adopt a hybrid strategy for edit-wise explanations, where each edit is explained through three aspects, including error type labels, error severity levels, and free-text error description. 1) **Error type labels** allow language learners to comprehend and infer syntax and grammar rules in an inductive manner. In particular, we employ a two-tier hierarchical taxonomy including 5 major types and 16 minor types shown in Table 1, inspired by existing studies (Liping, 2014; Peng et al., 2021; Zhang et al., 2022). The detailed description of various error types are included in Appendix A.1 and A.2. If an edit covers multiple error types, we select the one with the highest granule. 2) **Error severity levels**, ranging from 1 to 5 points, indicate the significance of a specific grammatical error. 3) **Error descriptions** are the most crucial and flexible element. These provide keywords, pertinent linguistic knowledge, causes of errors, and revision guidance in a free-text format. We stipulate well-defined error description should meet three principles: fluency, reasonability (making sense to humans), and

| Major Type | Minor Type |
|---|---|
| **Punctuation-level Error** | 标点冗余 (Punctuation Redundancy)<br>标点丢失 (Punctuation Missing)<br>标点误用 (Punctuation Misuse) |
| **Spelling-level Error** | 字音混淆错误<br>Phonetic Confusion Error<br>字形混淆错误<br>Glyph Confusion Error<br>词内部字符异位错误<br>Internal Character Misplacement Error<br>命名实体拼写错误<br>Named Entity Misspelling |
| **Word-level Error** | 词语冗余 (Word Redundancy)<br>词语丢失 (Word Missing)<br>词语误用 (Word Misuse) |
| **Sentence-level Error** | 词序不当 (Improper Word Order)<br>逻辑不通 (Illogicality)<br>句式杂糅 (Run-on Sentence) |
| **Other Special Error** | 照应错误 (Inconsistency Error)<br>歧义错误 (Ambiguity Error)<br>语气不协调 (Inconsistent Tone) |
| **Other** | |

Table 1: Hierarchical taxonomy of grammatical error types defined in our benchmark.

| Dataset | Sentences | Edits/Sent. | Chars/Sent. |
|---|---|---|---|
| **FCGEC** | 41,340 | 1.0 | 53.1 |
| **YACLC-minimal-dev** | 1,839 | 2.9 | 25.9 |
| **MuCGEC-dev** | 1,137 | 3.2 | 38.5 |
| **NaCGEC-dev** | 500 | 1.1 | 56.2 |
| **NLPCC-test** | 2,000 | 2.0 | 29.7 |
| **HSK** | 156,870 | 1.4 | 27.2 |
| **EXCGEC** (FCGEC) | 2,308 | 1.1 | 55.1 |
| **EXCGEC** (YACLC) | 1,235 | 3.5 | 24.3 |
| **EXCGEC** (MuCGEC-dev) | 789 | 3.3 | 40.4 |
| **EXCGEC** (NaCGEC-dev) | 449 | 1.1 | 56.1 |
| **EXCGEC** (NLPCC-test) | 1,611 | 1.7 | 28.9 |
| **EXCGEC** (HSK) | 1,824 | 2.1 | 32.0 |
| **EXCGEC** | 8,216 | 2.0 | 38.8 |

Table 2: Dataset statistics of the EXCGEC benchmark.

faithfulness (targeted to a specific edit). To ensure the reasonability and faithfulness, the error description must mostly conform to the syllogism form of deductive reasoning: *[major premise: semantic rules and related knowledge]*, *[minor premise: the reason for the error in the text]*, and *[explain how to correct it]*. Further, any evidence from the source $X$ must be enclosed within special markers 【】. Similarly, correction content that occurs in the target sentence $Y$ must be enclosed within { }, as indicated in Figure 1.

## 4.2 Explanation Synthesizing

Annotating high-quality explanations in a large scale poses a huge challenge to our benchmark construction. Hence, we leverage GPT-4 to synthesize edit-wise explanations efficiently. To achieve this, we first select 10,000 parallel samples across 6 existing benchmarks or datasets of Chinese GEC, with the details listed in Table 2. We pick out only the samples with changed target sentences, and select the single target sentence with the most edits if a sample is annotated with multiple target sentences. Then, we prompt GPT-4 to generate edit-wise explanations following in-context learning. To ensure faithfulness of synthesized explanation, we first extract edits using the toolkit CLEME (Ye et al., 2023). Inspired by Li et al. (2022), we then employ the Rationalization Prompting (RP) strategy, where we concatenate task definition, demonstrations, and a parallel sample $(X, Y)$ with extracted edits $E = \{e_1, e_2, \cdots, e_n\}$ as the prompt. For each error type, we provide the definition, a suggested template of error description, and a demonstration. The prompt is listed in Appendix A.3.

## 4.3 Explanation Refinement and Analysis

Benefiting from the extensive knowledge acquired during the large-scale pre-training process, GPT-4 is able to generate fluent, reasonable and plausible explanations in most cases, meeting the requirements with specified instructions. However, GPT-4 is not guaranteed to produce all high-quality explanations due to hallucination, and the patterns of those invalid explanations are referred to as failure modes. Therefore, we hire 12 native speakers, all of whom are graduated students, to screen out invalid explanations. We finally obtain 8,216 clean samples out of 10,000 samples. We further investigate the failure modes of invalid explanations generated by GPT-4, which is provided in Appendix A.4.

## 4.4 Automatic Metrics

Recent studies leverage human evaluation for evaluation of GEE due to the lack of enough annotated samples, posing a challenge for efficient development of EXGEC systems. In this paper, we introduce a comprehensive set of automatic metrics for both correction and explanation parts.

**Correction.** We employ CLEME and ChERRANT to evaluate the correction performance. Both are edit-based metrics that output P/R/$F_{0.5}$ scores, and they have been proven reliable metrics for GEC on CoNLL-2014 (Ye et al., 2023).

**Explanation.** Since an edit-wise explanation consists of three critical elements, we define respectively automatic metrics for them. 1) Accuracy

**Algorithm 1** COTE Decoding Algorithm

**Input:** Source text $X$, a post-explaining model $\mathcal{M}$, and the edit extraction function $f$.
**Output:** Target text $\hat{Y}$, and explanations $\hat{E}'$.

$\hat{Y} \leftarrow \text{BeamSearch}(\mathcal{M}(\text{Json}(X)))$
$\hat{E}' \leftarrow \emptyset$
**if** $\hat{Y} = X$ **then**
    **return** $\hat{Y}, \hat{E}'$
**end if**
$E \leftarrow f(X, \hat{Y})$
$\hat{E}' \leftarrow \text{Top-P}(\mathcal{M}(\text{Json}(X, Y, E)))$
**return** $\hat{Y}, \hat{E}'$

and Macro-F1 scores are computed for error type clarification, following the conventional evaluation protocol of text clarification (Li et al., 2020). 2) We report mean absolute error (MAE) to shown the deviation of hypothesis error severity levels towards ground truth ones. 3) We employ various metrics for evaluating the free-text explanation description, including BLEU, METEOR, ROUGE-1, ROUGE-2, and ROUGE-L. We leave the analysis on effectiveness of these metrics to Section 7.2.

# 5 Method

## 5.1 Training

To streamline the training process covering all the tasks mentioned in Section 3, we treat all of them as a unified Seq2Seq task. To achieve this, we linearize the data in the format of json (Gao et al., 2023). This structured approach simplifies the process of output parsing involving three types elements of edit-wise explanations, and provides a consistent and controllable view to distinguish tasks, enabling the model understand essential task elements and their relations. Therefore, we train all models using the same smooth cross entropy loss, regardless of the specific task.

## 5.2 Inference

For post-explaining EXGEC models, we design a specific **C**orrect-**T**hen-**E**xplain decoding algorithm called **COTE**, which is presented in Algorithm 1. First, we employ the greedy beam search decoding strategy for the correction part, which is beneficial to relieve the over-correction problem that is common on LLMs (Cao et al., 2023; Loem et al., 2023; Li et al., 2023a). Then, we apply CLEME to extract edits. Notably, we merge adjacent edits with distance less than 2 characters to avoid fragmented edits. Finally, we leverage the Top-p decoding strategy for generating explanations, encouraging diversified natural language explanations. It is worth noting that COTE is not accessible to pre-explaining models since the edit extraction tool necessitates both a source text and a target text.

# 6 Experiments

## 6.1 Experimental Settings

**Backbones.** We benchmark three series of LLMs, including Llama-3 (Touvron et al., 2023), Qwen-1.5 (Bai et al., 2023), and DeepSeek (Bi et al., 2024). For each series of LLMs, we experiment with their base and chat (or instruct) versions to investigate whether further alignment training benefits the task. All results are based on EXCGEC-test. Training details are reported in Appendix B.1.

**Evaluation.** We report experiment results using the metrics introduced in Section 4.4, calculated using open-source toolkits including *NLTK* (Bird and Loper, 2004), *rouge* (Lin, 2004), and *scikit-learn* (Pedregosa et al., 2011). Particularly, we observe many hypothesis edits are not covered in references, making it impossible to evaluate the subsequent explanations for these edits. To address this, we introduce two extra indicators, namely *Hit* and *Miss* rates. A hypothesis edit overlapping with a reference edit is designated as a hit edit, while a reference edit without any match with hypothesis edits is deemed a miss edit. The hit rate is defined as the ratio of hit edits to all hypothesis edits, and the miss rate as the ratio of miss edits to all reference edits. Only hit edits are used to determine the evaluation outcomes for explanations.

## 6.2 Results of Multi-task Models

The preliminary results from both the post- and pre-explaining models are presented in Table 3, from which we can make some conclusions.

**Post-explaining models consistently outperform pre-explaining models.** In relation to the correction aspect, all post-explaining models obtain higher $F_{0.5}$ scores than pre-explaining models, regardless of the applied backbones. A similar pattern is observed in the explanation part, where all the pre-explaining models invariably underperform their post-explaining counterparts. This suggests that a complexity for LLMs in initially explaining grammatical errors. And once pre-explaining models generate flawed explanations, the ensuing distraction impedes their ability to accurately correct the source text.

| Model | Correction↑ | | Explanation | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CLEME (P / R / $F_{0.5}$) | ChERRANT (P / R / $F_{0.5}$) | Hit↑ | Miss↓ | Acc↑ | F1↑ | MAE↓ | BLEU↑ | METEOR↑ | ROUGE- (1 / 2 / L)↑ |
| **Qwen1.5-7B-base** | 26.00 / 26.54 / **26.10** | 33.87 / **20.16** / 29.81 | 67.29 | 56.81 | 60.99 | **29.82** | 0.80 | 15.22 | **39.05** | 49.74 / 23.28 / 34.32 |
| **Qwen1.5-7B-chat** | **28.31** / 21.21 / **26.54** | **36.74** / 17.26 / **29.98** | 68.94 | 64.83 | **61.98** | 29.62 | **0.75** | **15.49** | 38.88 | **50.32 / 24.25 / 35.24** |
| **Llama3-8B-base** | 20.92 / 23.60 / 21.40 | 28.81 / 17.78 / 25.63 | 61.54 | 58.38 | 58.39 | 25.12 | 0.91 | 14.54 | 37.84 | 49.53 / 23.19 / 34.58 |
| **Llama3-8B-instruct** | 21.33 / 26.05 / 22.14 | 29.00 / 19.40 / 26.39 | 61.40 | **55.71** | 59.16 | 25.63 | 0.88 | 14.70 | 36.89 | 49.41 / 23.54 / 34.87 |
| **DeepSeek-7B-base** | 26.21 / 7.00 / 16.92 | 36.00 / 7.04 / 19.75 | **69.92** | 85.39 | 60.64 | 26.47 | 0.79 | 15.07 | 38.05 | 50.19 / 24.10 / 34.90 |
| **DeepSeek-7B-chat** | 25.46 / 18.51 / 23.68 | 34.02 / 15.75 / 27.62 | 67.52 | 66.64 | 58.11 | 24.45 | 0.84 | 13.94 | 36.97 | 48.66 / 22.70 / 34.23 |
| **Qwen1.5-7B-chat** | **13.76 / 13.42 / 13.69** | **19.27 / 9.93 / 16.22** | **29.49** | 80.24 | 23.35 | 8.22 | **1.17** | **7.75** | **27.67** | **40.47 / 15.00 / 28.20** |
| **Llama3-8B-instruct** | 7.12 / 11.17 / 7.68 | 10.86 / 8.57 / 10.31 | 23.88 | **73.06** | **24.31** | **8.78** | 1.21 | 5.78 | 23.07 | 37.57 / 13.47 / 27.19 |
| **DeepSeek-7B-chat** | 9.93 / 8.26 / 9.55 | 14.28 / 7.07 / 11.86 | 24.72 | 78.67 | 19.12 | 5.84 | 1.29 | 5.91 | 23.95 | 37.59 / 13.11 / 26.78 |

Table 3: Main results of multi-task learning models. Results of post-explaining models are listed in the *top* block, while those of pre-explaining models are in the *bottom* block.

| Model | Correction↑ | | Explanation | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CLEME (P / R / $F_{0.5}$) | ChERRANT (P / R / $F_{0.5}$) | Hit↑ | Miss↓ | Acc↑ | F1↑ | MAE↓ | BLEU↑ | METEOR↑ | ROUGE- (1 / 2 / L) |
| **Qwen1.5-7B-base** | — / — / — | — / — / — | 98.42 | 6.14 | 85.00 | **43.32** | 0.73 | 19.70 | 43.18 | 53.48 / 27.79 / 38.12 |
| **Qwen1.5-7B-chat** | 62.59 / 87.35 / 66.35 | 67.58 / 69.53 / 67.96 | **99.93** | 0.43 | 81.53 | 39.56 | 0.73 | 17.88 | 41.40 | 51.73 / 28.81 / 36.51 |
| **Llama3-8B-base** | — / — / — | — / — / — | 99.42 | 2.27 | 83.27 | 40.51 | 0.89 | 20.52 | 43.37 | 54.32 / 29.05 / 39.49 |
| **Llama3-8B-instruct** | **69.10 / 90.90 / 72.58** | **73.75 / 74.37 / 73.87** | 99.63 | 1.67 | **85.99** | 41.84 | 0.78 | 20.73 | 42.98 | **54.60 / 29.64 / 40.04** |
| **DeepSeek-7B-base** | — / — / — | — / — / — | **99.93** | 3.54 | 85.06 | 40.19 | **0.71** | **20.78** | **43.48** | 54.07 / 29.18 / 39.57 |
| **DeepSeek-7B-chat** | 41.12 / 79.02 / 45.48 | 48.35 / 53.20 / 49.25 | **99.93** | **0.40** | 81.17 | 35.93 | 0.74 | 19.57 | 42.32 | 53.12 / 28.03 / 38.59 |

Table 4: Ground truth results of multi-task learning models. We report the explanation performance (**right** block) of *post-explaining* models conditioned on source texts and ground truth target texts. Contrarily, we report the correction performance (**left** block) of *pre-explaining* models conditioned on source sentences and ground truth explanations.

| Model | Correction↑ | | Explanation | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CLEME (P / R / $F_{0.5}$) | ChERRANT (P / R / $F_{0.5}$) | Hit↑ | Miss↓ | Acc↑ | F1↑ | MAE↓ | BLEU↑ | METEOR↑ | ROUGE- (1 / 2 / L) |
| **Post-explaining** | 28.31 / 21.21 / 26.54 | 36.74 / 17.26 / 29.98 | 68.94 | 64.83 | 61.98 | 29.62 | **0.75** | 15.49 | 38.88 | 50.32 / 24.25 / 35.24 |
| **Pre-explaining** | 13.76 / 13.42 / 13.69 | 19.27 / 9.93 / 16.22 | 29.49 | 80.24 | 23.35 | 8.22 | 1.17 | 7.75 | 27.67 | 40.47 / 15.00 / 28.20 |
| **Pipeline** | **32.45 / 23.93 / 30.29** | **40.50 / 19.58 / 33.37** | **88.53** | **57.42** | **74.80** | **34.84** | **0.75** | **16.44** | **39.76** | **50.62 / 24.56 / 35.71** |

Table 5: Comparison of the multi-task solutions and the GEC-GEE pipeline solution based on Qwen1.5-7B-chat.

**Chat models outperform base models.** For post-explaining models, we observe all chat or instruct models gain slightly higher $F_{0.5}$ correction scores, and they also marginally outperform their base version counterparts in the explanation task. It indicates that additional alignment training (Wang et al., 2023) can benefit the EXGEC task.

### 6.3 Ground Truth Results

In order to study the isolated performance of multi-task models, we provide part ground truth information in advance during the inference stage. Specifically, we provide ground truth target texts for post-explaining and report their performance of explanation. Conversely, we offer ground truth explanations for pre-explaining and report their performance of correction. This experimental setting allows for revealing the specialized performance, eliminating the distraction of previously generated contents. The results are presented in Table 4.

**For the task of explanation, two base models slightly outperform chat models.** Specially, the base version models of Qwen and DeepSeek exhibit a minor increase in performance over their chat/instruct counterparts on classifying error types and providing error descriptions. However, this is not true for Llama3, where the Llama3-instruct model obtain the highest Acc, METEOR and ROUGE scores. Also noteworthy is the significantly lower miss rates of chat/instruct models compared to base models, indicating a tendency for the latter to overlook explanations, even when ground truth target texts are available. These findings contradict the joint results in Table 3. We speculate the reason is base models may be more susceptible to low-quality self-generated corrections.

**Ground truth explanations tremendously improve correction performance.** Since the explanations include explicit clues for corrections such as evidence words and revision advice, it is effort-

| Model | Correction↑ | | Explanation | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CLEME (P / R / $F_{0.5}$) | ChERRANT (P / R / $F_{0.5}$) | Hit↑ | Miss↓ | Acc↑ | F1↑ | MAE↓ | BLEU↑ | METEOR↑ | ROUGE- (1 / 2 / L) |
| **Beam search** | **28.31** / 21.21 / **26.54** | **36.74** / 17.26 / **29.98** | 99.22 | 19.05 | **83.93** | **44.48** | **0.71** | **22.71** | **44.28** | **55.55 / 32.26 / 42.34** |
| **Top-p** | 19.45 / **27.05** / 20.61 | 24.83 / **19.14** / 23.44 | **99.93** | **0.40** | 81.53 | 39.56 | 0.74 | 17.88 | 41.40 | 51.73 / 25.81 / 36.51 |

Table 6: Comparison of the post-explaining model with different token-wise decoding strategies. Note that the explanation performance is conditioned on ground truth target texts in order to exclude unrelated interference.

| | Hit↑ | Miss↓ | Acc↑ | F1↑ | MAE↓ | BLEU↑ | METE↑ | ROUGE- (1/2/L)↑ |
|---|---|---|---|---|---|---|---|---|
| **w COTE** | **99.93** | **0.43** | **81.53** | **39.56** | **0.74** | **17.88** | **41.40** | **51.73 / 25.81 / 36.51** |
| **w/o COTE** | 49.64 | 54.01 | 42.51 | 17.77 | 0.93 | 11.53 | 33.81 | 46.35 / 19.34 / 31.28 |

Table 7: Ablation results of COTE from the same Qwen1.5-7B-chat post-explaining model.

| | Pearson | Spearson |
|---|---|---|
| **Human** v.s. **BLEU** | 0.9222 | 0.6571 |
| **Human** v.s. **METEOR** | 0.9280 | 0.7714 |
| **Human** v.s. **ROUGE-1** | 0.9464 | 0.8286 |
| **Human** v.s. **ROUGE-2** | 0.9175 | 0.4857 |
| **Human** v.s. **ROUGE-L** | 0.9352 | 0.6571 |
| $\mathbf{A}_1$ v.s. $\mathbf{A}_2$ | 0.9874 | 0.9429 |

Table 8: Correlations between human judgements and metrics for error descriptions.

less for pre-explaining models to correct the source.

### 6.4 Comparison with Pipeline

We compare the results of multi-task models and GEC-GEE pipeline with COTE in Table 5. It indicates that the pipeline can improve both correction and explanation performance, highlighting the challenges of learning a multi-task model for EXGEC.

## 7 Analysis

### 7.1 Ablation Results

We conduct ablation studies on Qwen1.5-7B-chat to provide in-depth insights into post-explaining models. We also study the effect of model sizes in Appendix B.2 and provide a case study for different LLMs in Appendix B.3.

**Effect of COTE.** The impact of COTE introduced in Section 5.2 is examined in this section. We provide the post-explaining model with ground truth target texts, which allows us to focus on the explanation performance. The results presented in Table 7 reveal a huge performance drop if we do not leverage COTE, especially the hit and miss rates. This demonstrates the effectiveness of COTE.

**Effect of token-wise decoding strategies.** By default, we employ beam search decoding for corrections and top-p decoding for explanations. In this section, we explore the reverse setting, and the results are reported in Table 6. When switching from beam search to top-p for correction, we observe a huge performance drop in precision and $F_{0.5}$ and increase in recall, which means top-p encourages LLMs to over-correct (Cao et al., 2023). On the other hand, leveraging beam search improves explanation performance, suggesting the potential benefits of a greedy decoding algorithm for the task. However, we notice that beam search also increases the miss rate. We speculate that beam search may discard some low-likelihood explanations.

### 7.2 Human Evaluation for Error Descriptions

Despite the efficiency of automatic metrics in evaluating error descriptions, their accuracy remains to be confirmed. Therefore, this section attempts to demonstrate the suitability of different metrics by comparing their corrections with human judgements. We report the correlations between two human annotators and the ones between average human ratings and metric scores in Table 8. We observe the inter-annotator correlations are close to 1, meaning it is relatively easy to determine the quality of error descriptions for human. Among various metrics, ROUGE-1 achieve the highest correlations, followed by METEOR. All the introduced metrics show moderate or high correlations, indicating that it is advisable to employ them as proxies for human evaluation. We provide the detailed annotation guidance and rating rules in Appendix B.4.

## 8 Conclusion

We propose and formulate the task of EXGEC, overcoming the limitation of previous studies that fail to establish the interaction of both correction and explanation tasks. To develop the task, we propose the EXCGEC benchmark, based on which we develop baseline models in multiple settings. Extensive experiments and analyses reveal several challenges of the task, and we hope this paper can serve as a starting point for future exploration.

## Limitations

**Inferior performance of multi-task models.** In our experiments, we observe the pipeline solution outperform the multi-task solutions, regardless of correction or explanation tasks. This suggests that the multi-task models struggle to reap positive benefits from the interaction of both tasks. We leave the exploration of effective multi-task learning EXGEC models to the future work.

**Limitations of synthesizing datasets.** LLM-augmented datasets may include some unintended biases towards or inaccuracies, resulting in skewed or unfair outcomes in applications. Second, it is necessary to manually filter out invalid data in order to ensure the quality of datasets. But it is indeed a advisable method to construct datasets using LLMs, considering its efficiency.

**Adaptation to other languages.** The general design of our proposed edit-wise explanations can be easily adapted to other languages. However, the detailed design may not be suitable to other languages. For example, the two-tier hierarchical taxonomy of error types is tailored for Chinese.

## Ethics Statement

Our proposed benchmark is built upon existing datasets, backbones and metrics, all of which are publicly available. We have cited the corresponding authors or projects of them, and confirm that they are consistent with their intended use.

Additionally, we conduct human evaluation experiments to ensure the quality of the dataset and find out the correlations between metrics and human judgements. To achieve this, we hire 12 native speakers, all of whom are graduated students. Each annotator could complete the entire annotation process within approximately 6∼8 working hours. All annotators were paid for their work, with an average salary of approximately $5 per hour.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Gloria Ashiya Katuka, Alexander Gain, and Yen-Yun Yu. 2024. Investigating automatic scoring and feedback using large language models. *arXiv e-prints*, pages arXiv–2405.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, Øistein E Andersen, and Ted Briscoe. 2019a. The bea-2019 shared task on grammatical error correction. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 52–75.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019b. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

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

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3):643–701.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Qi Cao, Takeshi Kojima, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Unnatural error correction: Gpt-4 can almost perfectly handle unnatural scrambled text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8898–8913.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. *arXiv preprint arXiv:2403.02990*.

Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. 2023. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33.

Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. 2023. Enhancing grammatical error correction systems with explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7489–7501, Toronto, Canada. Association for Computational Linguistics.

Mariano Felice, Christopher Bryant, and Ted Briscoe. 2016. Automatic extraction of learner errors in esl sentences using linguistically enhanced alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 825–835.

Chang Gao, Wenxuan Zhang, Guizhen Chen, and Wai Lam. 2023. Jsontuning: Towards generalizable, robust, and controllable instruction tuning. *arXiv preprint arXiv:2310.02953*.

Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui. 2021. Exploring methods for generating feedback comments for writing learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9719–9730.

Mareike Hartmann and Daniel Sonntag. 2022. A survey on improving NLP models with human explanations. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 40–47, Dublin, Ireland. Association for Computational Linguistics.

Xuanli He, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. 2023. Using natural language explanations to improve robustness of in-context learning for natural language inference. *arXiv preprint arXiv:2311.07556*.

Masahiro Kaneko and Naoaki Okazaki. 2023. Controlled generation with prompt insertion for natural language explanations in grammatical error correction. *arXiv preprint arXiv:2309.11439*.

Masahiro Kaneko, Sho Takase, Ayana Niwa, and Naoaki Okazaki. 2022. Interpretability for language learners using example-based grammatical error correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7176–7187, Dublin, Ireland. Association for Computational Linguistics.

Yunshi Lan, Xinyuan Li, Hanyue Du, Xuesong Lu, Ming Gao, Weining Qian, and Aoying Zhou. 2024. Survey of natural language processing for education: Taxonomy, systematic review, and future trends. *arXiv preprint arXiv:2401.07518*.

Dongfang Li, Baotian Hu, Qingcai Chen, Tujie Xu, Jingcong Tao, and Yunan Zhang. 2022. Unifying model explainability and robustness for joint text classification and rationale extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10947–10955.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. 2020. A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*.

Shiyang Li, Jianshu Chen, yelong shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2024. Explanations from large language models make small reasoners better. In *2nd Workshop on Sustainable AI*.

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

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiang Liping. 2014. *HSK Standard Course 1 (Chinese and English Edition)*. Beijing Language & Culture University Press,China; First Edition.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of gpt-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. *arXiv preprint arXiv:2305.18156*.

Marta Montenegro-Rueda, José Fernández-Cerero, José María Fernández-Batanero, and Eloy López-Meneses. 2023. Impact of the implementation of chatgpt in education: A systematic review. *Computers*, 12(8):153.

Ryo Nagata. 2019. Toward a task of feedback comment generation for writing learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3206–3215.

Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, Masato Mita, Artem Chernodub, and Olena Nahorna. 2021. Shared task on feedback comment generation for language learners. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 320–324.

Ryo Nagata, Kentaro Inui, and Shin'ichiro Ishikawa. 2020. Creating corpora for research in feedback comment generation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 340–345.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Yue Peng, Wei Yan, and Liying Cheng. 2021. Hanyu shuiping kaoshi (hsk): A multi-level, multi-purpose proficiency test. *Language Testing*, 38(2):326–337.

Kashun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12113–12139, Singapore. Association for Computational Linguistics.

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

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint arXiv:2404.15845*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yingying Wang, Cunliang Kong, Liner Yang, Yijun Wang, Xiaorong Lu, Renfen Hu, Shan He, Zhenghao Liu, Yun Chen, Erhong Yang, et al. 2021. Yaclc: a chinese learner corpus with multidimensional annotation. *arXiv preprint arXiv:2112.15043*.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. LLM-powered data augmentation for enhanced cross-lingual performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686, Singapore. Association for Computational Linguistics.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

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

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. *arXiv preprint arXiv:2204.10994*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

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 7*, pages 439–445. Springer.

# A Benchmark Details

## A.1 Description of Grammatical Error Types

In the taxonomy of Chinese grammatical errors, we first divide grammatical errors by their effecting granular into 5 major types, including punctuation-level, spelling-level, word-level, sentence-level, and other special errors. This section will clarify all involved grammatical error types by providing detailed description and corresponding examples listed in Table ??? .

**Punctuation-level Error.** This type of grammatical errors primarily involves redundancy, missing, and misuse of punctuation.

- 标点冗余 (**Punctuation Redundancy**). The type of punctuation redundancy refers to the insertion of punctuation in unnecessary places. For punctuation redundancy errors, first explain the role of the punctuation symbols involved, and then explain the reasons for punctuation redundancy in the current case.

- 标点丢失 (**Punctuation Missing**). Punctuation missing mainly refers to the omission of punctuation that should have existed in the middle and end of a sentence. For the explanation of punctuation missing errors, first point out the evidence words and missing punctuation symbols, and then explain the role of the added punctuation in this context.

- 标点误用 (**Punctuation Misuse**). Misuse of punctuation is very common in daily Chinese writing. For cases of punctuation misuse, first briefly explain the roles of misused punctuation and correct punctuation, and then explain sufficient reasons for correction.

**Spelling-level Error.** Spelling-level errors refer to people who, due to carelessness or lack of knowledge, write incorrect characters or words during the writing process. The type is so common that Chinese Spelling Check (CSC), as a standard NLP task specialized in spelling-level errors, attract the attention from many researchers. Inspired by these studies, we categorize spelling-level errors further into 4 sub-classes.

- 字音混淆错误 (**Phonetic Confusion Error**). Phonetic confusion errors are caused by misusing the Chinese characters with the same or similar pinyin. The vast majority of Chinese Internet users are using pinyin input method, so many Chinese spelling-level errors on Internet fall in this type.

- 字形混淆错误 (**Glyph Confusion Error**). In addition to pinyin input method, some users apply Wubi input method or other glyph-based input methods. In this case, they are prone to spelling errors due to confusion of fonts or strokes.

- 词内部字符异位错误 (**Internal Character Misplacement Error**). Internal character misplacement error refers to expressing a multi-character word in disorder of characters. The type seldom happens for native speakers, but sometimes in texts written by L2-speakers. For example, the spelling-level error "共公" falls in this type and should be corrected to "公共".

- 命名实体拼写错误 (**Named Entity Misspelling**). There are numerous named entity words in Chinese, such as person names, organization names, place names, and all other entities identified by terminologies. These words are also very prone to spelling errors.

**Word-level Error.** Word-level errors often refer to misuse of individual words or idioms in a sentence, but the syntactic structure of the sentence is correct. This type of error belongs to the most common category in Chinese text errors and can usually be subdivided into the following three types:

- 词语冗余 (**Word Redundancy**). The simultaneous appearance of words with the same or similar meanings in a sentence can cause semantic repetition and sentence redundancy, which is known as word redundancy. Repeated words often appear adjacent to each other, so it is important to pay attention to whether the meanings of adjacent words are exactly the same. If they are the same, it may lead to the problem of word redundancy.

- 词语丢失 (**Word Missing**). In modern Chinese, sentences generally have six major components, namely subject, predicate, object, attributive, adverbial, complement, etc. A sentence must express a complete meaning, and its structure must also be complete. The so-called complete structure does not mean that a sentence must have the usual six components, but rather that the sentence should be composed of the necessary components to express the complete meaning. If the necessary sentence components are missing, it will cause the phenomenon of word missing.

- 词语误用 (**Word Misuse**). Word Misuse indicates improper use of words in the text. The main cause of this error is the author's insufficient understanding of the meaning and part of speech of a certain word.

**Sentence-level Error.** This type mainly involves sentence-level issues, not just individual words or characters. Sentence-level errors are often caused by violating common syntactic structures, or not following objective reasoning.

- 词序不当 (**Improper Word Order**). Proper word order is essential to express exact meaning in Chinese. Writing texts without accurate word order results in the type of improper word order. If a sentence is not combined according to the intended meaning, it may lead to confusion in the sentence structure, resulting in an imbalance in the relationship between sentence components and affecting the expression of sentence meaning.

- 逻辑不通 (**Illogicality**). Illogicality refers to a sentence that conforms to grammatical norms but does not conform to logical reasoning. Illogicality can be caused by many reasons such as improper logical order, causal confusion, and reversal of subject and object.

- 句式杂糅 (**Run-on Sentence**). Run-on Sentence in Chinese usually refers to the use of two formats or sentences with similar or identical meanings in one sentence. People originally used one format when writing sentences, but due to interference from other factors such as sentence content, they may unconsciously switch to another format, resulting in a mixture of the two formats.

**Other Special Error.** Besides the above grammatical error types, other several types can not easily fit in the mentioned major types. So we classify them to other special errors.

- 照应错误 (**Inconsistency Error**). Inconsistency errors are ones involved in the mistaken referential relationship between two words, and explaining this grammatical error requires knowledge of the referential relationship between each word.

- 歧义错误 (**Ambiguity Error**). Ambiguity errors happen when a word or a sentence can be understood as having multiple meanings.

- 语气不协调 (**Inconsistent Tone**). Inconsistent tone refers to the inconsistency of tone between the preceding and following sentences.

Additionally, we define the grammatical error type *Other* as ones that do not fit in any of the above error types. These errors are usually involved in rather significant modification and sometimes change the original semantics.

### A.2 Examples of Error Types

We list the examples of error types in Figure 4, 5, 6.

### A.3 Prompt of Generating Explanations

The prompt we use to generate explanations is shown in Figure 8. We also provide an English version in Figure 9.

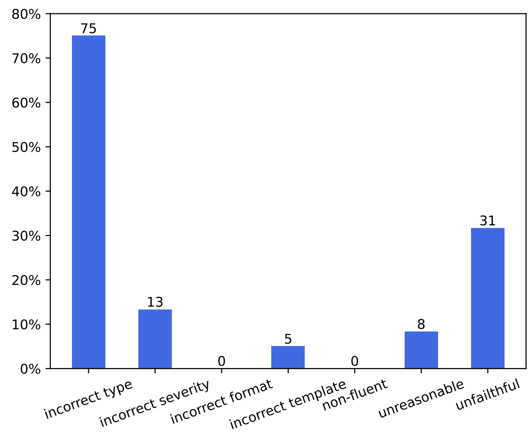


Figure 3: Distribution of 7 kinds of LLM errors.

| Configuration | Value |
|---|---|
| **Fine-tuning** | |
| Devices | 2 Tesla A100 GPU (80GB) |
| Epochs | 5 |
| Finetuning type | Lora |
| Train batch size per GPU | 2 |
| Eval batch size per GPU | 1 |
| Gradient accumulation steps | 16 |
| Optimizer | AdamW ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-6}$) |
| Learning rate | $5 \times 10^{-5}$ |
| Learning rate schedule | cosine decay |
| Warmup steps | 20 |
| Eval steps | 200 |
| Cutoff length | 1024 |
| Preprocessing workers number | 16 |
| Numerical precision | fp16 |
| Weight decay | 0.05 |
| **Inference** | |
| Beam size | 5 |
| Top-p | 0.8 |
| Max new tokens | 2048 |
| Temperature | 0.7 |

Table 9: Hyper-parameter values used in our experiments.

### A.4 Detailed Description of LLM Failure Modes

We categorize the failure modes in our case into seven major reasons: incorrect type, incorrect severity, incorrect format, incorrect template, non-fluency, unreasonability, and unfaithfulness. One expert annotator is asked to classified the sampled 100 invalid explanations, where an explanation may be categorized into multiple failure modes. The annotation results, illustrated in Figure 3, reveal that GPT-4 tend to mis-classify grammatical errors and providing unfaithful error descriptions. On the other hand, GPT-4 is capable to a large extend to offer well formed, fluent, and reasonable explanations, demonstrating the effectiveness of LLM annotation on this task.

The definitions of seven failure modes of expla-

| Model | Correction↑ | | Explanation | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CLEME (P / R / $F_{0.5}$) | ChERRANT (P / R / $F_{0.5}$) | Hit↑ | Miss↓ | Acc↑ | F1↑ | MAE↓ | BLEU↑ | METEOR↑ | ROUGE- (1 / 2 / L) |
| **Qwen1.5-1.8B-chat** | 21.11 / 19.28 / 20.72 | 28.91 / 15.70 / 24.74 | 59.94 | 65.14 | 55.80 | 23.27 | 0.89 | 10.19 | 34.35 | 48.66 / 22.70 / 34.23 |
| **Qwen1.5-4B-chat** | 22.49 / 20.84 / 22.14 | 30.57 / 16.85 / 26.29 | 62.91 | **62.70** | 57.16 | 25.31 | 0.85 | 11.61 | 35.91 | 46.83 / 19.59 / 30.86 |
| **Qwen1.5-7B-chat** | **28.31 / 21.21 / 26.54** | **36.74 / 17.26 / 29.98** | **68.94** | 64.83 | **61.98** | **29.62** | **0.75** | **15.49** | **38.88** | **50.32 / 24.25 / 35.24** |

Table 10: Comparison of post-explaining models with various model sizes.

nations are as follows:

- **Incorrect type**: the error type is incorrect.
- **Incorrect format**: the evidence content and the correction content are not highlighted by special markers 【 】 or { }.
- **Incorrect template**: the error description does not follow the syllogism form of deductive reasoning.
- **Non-fluency**: the error description is non-fluent or unreadable.
- **Unreasonability**: the error description contains obvious mistakes about linguistics, thus making it unacceptable for human.
- **Unfaithfulness**: the error description is not targeted to the given edit.

# B Experimental Details and Extra Results

## B.1 Implementation Details.

We train all models for 5 epochs and select the best model validated on EXCGEC-dev and report its performance on EXCGEC-test. The detailed training hyperparameter values of the all models in our experiments are shown in Table 9.

## B.2 Effect of Model Sizes

Table 10 indicates the varying performance across model sizes ranging from 1.8B to 7B. We observe consistent performance enhancement with increasing model sizes.

## B.3 Case Study

We provide a case study in Table 7.

## B.4 Details of Human Rating

Specifically, we hire 2 native Chinese speakers to rate the explanations generated by 6 post-explaining models in Table 3 conditioned on ground truth target texts. The rating scores range from 0 to 100, and each annotator concurrently rate 6 explanations for each sample. We randomly select 100 samples for annotation. We provide annotators with general scoring suggestions:

- **100 points**: Explain and describe fluently (fluency), introduce relevant semantic knowledge to enhance persuasiveness (rationality), and explain that it is aimed at the current editor (loyalty). All aspects are impeccable, and there is almost no better explanation or description than this.
- **80∼100 points**: The explanation and description are expressed fluently, satisfy fidelity, and have a certain degree of rationality, but there are certain degrees of flaws.
- **60∼80 points**: The explanation and description are expressed fluently, but the fidelity or rationality is not good enough, but it is somewhat helpful for correcting the grammar error in understanding.
- **30∼60 points**: The explanation and description are expressed fluently, but the rationality is poor, and it is not very helpful for correcting the grammar error in understanding.
- **0∼30 points**: The explanation and description are expressed fluently, but the fidelity is poor, and the object of explanation is not the current editor. There is no help in correcting the grammar error for understanding.
- **0∼30 points**: The explanation and description are vague and cannot be understood. There is no help in correcting the grammar error for understanding.

```
1 标点级别错误
# 标点冗余
{
   "input": "所以一些人说，：“读书一点用处都没有。”"
   "output": "所以一些人说：“读书一点用处都没有。”",
      "explanations": [
      {
         "error_severity": 1,
         "error_type": "标点冗余",
         "error_description": "【：】直接用于【说】等总说性或提示性词语后面，提起下文，没有必要在【：】前插入逗号。应删去【说】之
         后的冒号。"
      }
   ]
},

# 标点丢失
{
   "input": "人为了生存不管是干净的空气还是污染的空气都要呼吸。"
   "output": "人为了生存，不管是干净的空气还是污染的空气，都要呼吸。",
      "explanations": [
      {
         "error_severity": 1,
         "error_type": "标点丢失",
         "error_description": "【人为了生存】和【不管……】是两个分句，复句内各分句之间应使用逗号表示停顿。应在【为了生存】后添加逗
         号。"
      }
   ]
}


# 标点误用
{
   "input": "那我们一定要参加这个活动吗。"
   "output": "那我们一定要参加这个活动吗？",
      "explanations": [
      {
         "error_severity": 1,
         "error_type": "标点误用",
         “error_description”: "句号主要表示句子的陈述语气，而问号主要表示句子的疑问语气。【吗】意味着该句是一个疑问句，故应【吗】
         后的句号改为问号。"
      }
   ]
}

2 拼写级别错误
# 字音混淆错误
{
   "input": "我们舒舍有四个人。"
   "output": "我们宿舍有四个人。",
      "explanations": [
      {
         "error_severity": 1,
         “error_type”: "字音混淆错误",
         “error_description”: "{宿舍}指学校或用人单位等提供给学生和职工的房屋，对应句子中的【有四个人】。{宿舍}和【舒舍】发音相
         近，导致了此处的拼写错误。应将【舒舍】改为{宿舍}。"
      }
   ]
}

# 字形混淆错误
{
   "input": "这座关溢非常雄伟。"
   "output": "这座关隘非常雄伟。",
      "explanations": [
      {
         "error_severity": 1,
         “error_type”: "字形混淆错误",
         “error_description”: "{关隘}指险要的关口，在交通要道设立的防务设施，又称关卡。{隘}和【溢】字形相近，导致了此处的拼写错
         误。应将【关溢】改为{关隘}。"
      }
   ]
}

# 词内部字符异位
{
   "input": "我非常爱吃阴冬功。"
   "output": "我非常爱吃冬阴功。",
      "explanations": [
      {
         "error_severity": 2,
         "error_type": "词内部字符异位",
         “error_description”: "{冬阴功}是泰国和老挝的一道富有特色的酸辣口味汤品，书写者错误地将该词写成{阴冬功}。应将【阴冬功】
         改为{冬阴功}。"
      }
   ]
}
```

Figure 4: Examples of error types.

```
# 命名实体拼写错误
{
  "input": "我们都是海南詹州人。"
  "output": "我们都是海南儋州人。",
    "explanations": [
    {
      "error_severity": 2,
      “error_type”: “命名实体拼写错误",
      "error_description": "中国【海南】不存在【詹州】这一地名，但存在字形相近的{儋州}。【詹】与{儋}字形相近，导致了此处的拼
      写错误。应将【詹州】改为{儋州}。"
    }
  ]
},


3 词语级别错误
# 词语冗余
{
  "input": "终于看到了大熊猫，儿子显得特别兴奋极了。"
  "output": "终于看到了大熊猫，儿子显得特别兴奋。",
    "explanations": [
    {
      "error_severity": 3,
      “error_type”: “词语冗余",
      "error_description": "【特别】与【极了】都是修饰【兴奋】的程度副词，两者重复。应删去【特别】与【极了】其中一个。"
    }
  ]
},

# 词语丢失
{
  "input": "最终经过他的不懈努力，成为了一个地位很高的长官。"
  "output": "最终经过不懈努力，他成为了一个地位很高的长官。",
    "explanations": [
    {
      "error_severity": 3,
      “error_type”: “词语丢失",
      "error_description": "状语从句【经过不懈努力】和主句的谓语【成为】具有共同主语【他】，此处将【他】放在【经过】的后面导致
      句子缺失主语。可以把【他】放在【成为】之前，也可以把【他】提到【经过】的前面，充当状语从句和主句的共同主语。"
    }
  ]
}


# 词语误用
{
  "input": "这样一个年过八旬的老奶奶在她即将逝去的生命中仍然绽放着希望的光辉。"
  "output": "这样一个年过八旬的老奶奶在她即将逝去的生命中仍然散发着希望的光辉。",
    "explanations": [
    {
      "error_severity": 3,
      “error_type": “词语误用",
      “error_description": "谓语动词【绽放】和宾语【光辉】搭配不当，【绽放】一般用于形容花开时由花蕾花瓣紧闭展开的样子。应将
      【绽放】改为{散发}。"
    }
  ]
}

4 句法级别错误
# 词序不当
{
  "input": "改革开放后，中国的经济增长速度加快明显起来。"
  "output": "改革开放后，中国的经济增长速度明显加快起来。",
    "explanations": [
    {
      "error_severity": 4,
      “error_type": “词序不当",
      “error_description”: "状语【明显】用于修饰谓语动词【加快】，一般放在谓语动词之前。应将【明显】提到【加快】前面。""
    }
  ]
}

# 逻辑不通
{
  "input": "我们要注意多多提高总结自己。"
  "output": "我们要注意多多总结提高自己。",
    "explanations": [
    {
      "error_severity": 3,
      “error_type": “逻辑不通",
      “error_description”: "按照动作的发生顺序，应该先【总结】，再【提高】。应将逻辑顺序不当的【提高总结】改为{总结提高}。"
    }
  ]
}
```

Figure 5: Examples of error types.

# 句式杂糅
```
{
   "input": "形成沼泽的原因是水体沼泽化的结果。"
   "output": "形成沼泽是水体沼泽化的结果。",
      "explanations": [
      {
         "error_severity": 4,
         "error_type": "句式杂糅",
         "error_description": "【原因是……】和【是……的结果】都是表示原因的句式，将两个意思相同或相近的格式放在一个句子里使用导致
         格式杂糅。应删去其中一个。"
      }
   ]
}
```

5 其它特殊错误

# 照应错误
```
{
   "input": "蜗牛感到自己很悲哀，不只是因为它生命的长短。"
   "output": "蜗牛感到自己很悲哀，不只是因为它生命很短。",
      "explanations": [
      {
         "error_severity": 4,
         "error_type": "照应错误",
         "error_description": "生命的【短】可以照应【悲哀】，而【长】则不能照应。应将【生命的长短】改为{生命很短}。"
      }
   ]
},
```

# 歧义错误
```
{
   "input": "他没有听从导师的建议，对论文的观点加以修改，影响了论文的水平。"
   "output": "他没有根据导师的建议对论文的观点加以修改，影响了论文的水平。",
      "explanations": [
      {
         "error_severity": 4,
         "error_type": "歧义错误",
         "error_description": "句子有歧义，一方面可理解为导师建议他对论文的观点加以修改，但是他没有听从；另一方面可理解为导师认为
         论文不需要修改。本句可改为{他没有根据导师的建议对论文的观点加以修改，影响了论文的水平}。"
      }
   ]
}
```

# 语气不协调
```
{
   "input": "生活在这个万物滋长的社会中，我不禁慨叹社会的安详宁静仅仅是因为制度的限制吗？"
   "output": "生活在这个万物滋长的社会中，我不禁疑惑社会的安详宁静仅仅是因为制度的限制吗？",
      "explanations": [
      {
         "error_severity": 3,
         "error_type": "语气不协调",
         "error_description": "动词【慨叹】指感慨叹息，一般不用于疑问句。应将【慨叹】改为常用于疑问句的{疑惑}。"
      }
   ]
}
```

Figure 6: Examples of error types.

```
{
   "input": "我不想反叛宋老师的期待。"
   "output": "我不想辜负宋老师的期待。",

   # qwen1.5-7B-Chat
   "explanations": [
      {
         "error_type": "词语误用",
         "error_severity": 3,
         "error_description": "【反叛】通常指违反现有规则、制度或法律的行为，而根据上下文，"辜负期待"更符合句意，即没有达到或辜负期望。应将【反叛】改为{辜负}，以表达不辜负老师期望的意思。"
      }
   ],

   # qwen1.5-7B-Base
   "explanations": [
      {
         "error_type": "词语误用",
         "error_severity": 2,
         "error_description": "动词【反叛】一般指公开或秘密地反对政府或现有社会秩序的行为，而在这里应当使用表示辜负或使失望的动词。因此，应将【反叛】替换为{辜负}，以更准确地表达不想让宋老师失望的意思。"
      }
   ],

   # Llama3-Instruct
   "explanations": [
      {
         "error_type": "词语误用",
         "error_severity": 3,
         "error_description": "【反叛】通常指对某个制度、组织或权威的公开抵制或反对行为，而在这里应表达的是不想让宋老师失望的意思。正确的表达是【辜负】，表示不想让别人失望或不值得信赖。应将【反叛】改为{辜负}。"
      }
   ],

   # Llama3-Base
   "explanations": [
      {
         "error_type": "词语误用",
         "error_severity": 2,
         "error_description": "【反叛】一词通常指对某人或某个集体的背叛或叛变，而在此句中，作者表达的是对老师的期待的不负责任的含义，应该使用{辜负}。【辜负】一词含义是指不尽承诺或期望的义务，适用于此处语境。应将【反叛】替换为{辜负}。"
      }
   ],

   # DeepSeek-Chat
   "explanations": [
      {
         "error_type": "词语误用",
         "error_severity": 3,
         "error_description": "在此上下文中，【反叛】一词通常用于表达与权威或传统对立的行为，而【辜负】则用于表达未能达到他人的期望或未能履行承诺。根据句意，应表达的是未能满足老师期望的情感，所以应使用{辜负}。"
      }
   ],

   # DeepSeek-Base
   "explanations": [
      {
         "error_type": "词语误用",
         "error_severity": 2,
         "error_description": "在此上下文中，【反叛】通常用于表达对权威或传统观念的反抗或挑战，而【辜负】则表示没有达到或未能实现他人的期望。根据句意，{反叛}一词并不恰当，因为这里表达的是对宋老师期待的尊重和感激。应将【反叛】改为{辜负}，以更准确地表达对宋老师期待的尊重和感激之情。"
      }
   ],

   # GPT-4
   "explanations": [
      {
         "error_type": "词语误用",
         "error_severity": 3,
         "error_description": "动词【反叛】通常用来描述违背或背叛权威、组织或原则等的行为，而在这个语境中，作者意图表达的是不想让宋老师失望。因此，应用表达失望或希望落空的词汇{辜负}替换不当用词【反叛】，使句子意义正确传达。"
      }
   ]

}
```

Figure 7: A case study of all the LLMs involved in our experiments.

你是一个优秀的语法纠错解释模型，你需要在仔细对比错误句和纠正句的前提上，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、错误程度和错误描述。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。

每个语法错误由一个编辑改动（edit）来表示，为了提升解释的合理性和忠实性，你必须遵守以下原则：

1）必须对每个给定的语法错误进行解释，禁止私自修改编辑中的错误内容（src_content）和纠正内容（tgt_content）。

2）必须对每个语法错误分别给出相应的错误类型（error_type）、错误程度（error_severity）和错误描述（error_description）。

3）如果一处编辑改动存在多个语法错误，选择优先级最高的语法错误进行解释，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。

4）错误类型禁止自主捏造，只能来自下列错误类型：

- 标点冗余、标点丢失、标点误用
- 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
- 词语冗余、词语丢失、词语误用
- 词序不当、逻辑不通、句式杂糅
- 照应错误、歧义错误、语气不协调
- 其他错误

中的一个。语法错误类型将在下文给出定义和示例。当无法确定具体的错误类型时，统一分类为为“其他错误”。

5）错误程度的打分范围为1-5分，下面是每种分数在语法、语义层面上的详细描述和例句：

- 1分（无关紧要的错误）：可能是一些常规的打字错误或者一些影响很小的误用词语。例如：“他擅长数学和英语”应为“他擅长数学和英文”。
- 2分（轻度语法错误）：可能引起表达混淆，但并不会影响完整的理解。例如：“我喜欢狗和猫播放电子游戏”应为“我喜欢玩电子游戏，还喜欢狗和猫”。
- 3分（中度语法错误）：可能会导致句子部分不流畅，使读者需要重新阅读以理解含义。例如：“我走家去了”应为“我走去家了”。
- 4分（严重语法错误）：不仅会对理解产生困扰，还可能完全改变句子的意思。例如：“我想借用你的手机扮演职业摄影师”应为“我想借用你的手机拍摄一些专业的照片”。
- 5分（极度严重的语法错误）：可能导致句子无法理解。例如：“他举妈妈，我去购物车”应为“他举着妈妈的购物车，我就去了”。

6）错误描述必须符合演绎推理的三段论形式：[大前提：语义规则和相关知识] [小前提：当前文本的错误原因] [阐述如何纠正]

7) 错误描述需要提供充分且全面的纠正证据词，并使用以下符号强调错误描述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。

注意：下列大多数示例仅包含一个语法错误，但是正式输入数据通常包含多个语法错误，你必须对每个语法错误都分别给出相应的解释。输出必须严格符合json格式。

1 标点级别错误。即涉及标点符号的语法错误。

1.1 标点冗余：指在不必要的地方插入了标点。对于标点冗余错误，首先阐述所涉及标点符号的作用，然后解释标点冗余的原因。

解释标点冗余的建议模板为：[解释冗余标点和相关证据词的基本用法] [解释标点冗余的原因] 应删去[冗余标点]

标点冗余输入示例：

```
{
    "error_sentence": "所以一些人说，：“读书一点用处都没有。””,
    "correct_sentence": "所以一些人说：“读书一点用处都没有。””,
    "edit": [
        {
            "src_interval": [6,7],
            "tgt_interval": [6,6],
            "src_content": "，",
            "tgt_content": ""
        }
    ]
}
```

标点冗余输出示例：

```
{
     "edits": [
        {
            "src_interval": [6,7],
            "tgt_interval": [6,6],
            "src_content": "，",
            "tgt_content": "",
            "error_severity": 1,
            "error_type": "标点冗余",
            "error_description": "【：】直接用于【说】等总说性或提示性词语后面，提起下文，没有必要在【：】前插入逗号。应删去【说】之后的冒号。"
        }
    ]
}
```

1.2 标点丢失：主要指的是在句中、句末漏写了本应存在的标点。对于标点丢失错误的解释，首先要点明证据词和缺失的标点符号，然后阐述所加标点在此处的作用。

解释标点丢失的建议模板为：[解释丢失标点和相关证据词的基本用法] [解释标点丢失的原因] 应在[证据词]前/后添加[丢失标点]

……

Figure 8: The prompt used for explanation generation. For each error type, We provide the definition, a suggested template of error description, and a demonstration for GPT-4.

You are an excellent grammar error correction explanation model. Your task is to provide fluent, reasonable, and faithful explanations for punctuation errors, spelling errors, word errors, and syntactic errors in erroneous sentences by carefully comparing the erroneous sentences with the corrected sentences. The explanations should include the type of grammatical error, the severity of the error, and a description of the error. Fluency requires that the explanation itself has no grammatical errors and is expressed fluently; reasonableness requires that the explanation of the grammatical error is acceptable to people; faithfulness requires that all grammatical errors in the sentence have corresponding explanations, and the explanations should correspond to the correction methods of the correct sentence.

Each grammatical error is represented by an edit. To improve the reasonableness and faithfulness of the explanations, you must follow these principles:

1. Each given grammatical error must be explained, and the error content and correction content in the edits must not be modified.
2. Each grammatical error must be given a corresponding error type, error severity, and error description.
3. If an edit contains multiple grammatical errors, choose the grammatical error with the highest priority to explain. The priority order is: syntactic-level errors > word-level errors > spelling-level errors > punctuation-level errors.
4. Error types must not be fabricated; they can only come from the following error types:
    - Punctuation Redundancy, Punctuation Missing, Punctuation Misuse
    - Phonetic Confusion Error, Glyph Confusion Error, Internal Character Misplacement Error, Named Entity Misspelling
    - Word Redundancy, Word Missing, Word Misuse
    - Improper Word Order, Illogicality, Run-on Sentence
    - Inconsistency Error, Ambiguity Error, Inconsistent Tone
    - Other errors

    The definitions and examples of grammatical error types will be provided later. When it is impossible to determine the specific error type, classify it as "Other errors ".
5. The scoring range for error severity is 1-5 points. Here is a detailed description and examples of each score at the grammatical and semantic levels:
    - 1 point (trivial error): It may be some routine typing errors or minor word misuse that has little impact. Example: "他擅长数学和英语" should be "他擅长数学和英文".
    - 2 points (minor grammatical error): It may cause confusion in expression but does not affect the overall understanding. Example: "我喜欢狗和猫播放电子游戏" should be "我喜欢玩电子游戏，还喜欢狗和猫".
    - 3 points (moderate grammatical error): It may cause parts of the sentence to be incoherent, requiring the reader to reread to understand the meaning. Example: "我走家去了" should be "我走去家了".
    - 4 points (serious grammatical error): It not only causes confusion in understanding but may also completely change the meaning of the sentence. Example: "我想借用你的手机扮演职业摄影师" should be "我想借用你的手机拍摄一些专业的照片".
    - 5 points (extremely serious grammatical error): It may make the sentence incomprehensible. Example: "他举妈妈，我去购物车" should be "他举着妈妈的购物车，我就去了".
6. The error description must follow the deductive reasoning form of a syllogism: [Major premise: semantic rules and related knowledge] [Minor premise: the reason for the current text error] [Explain how to correct it].
7. The error description needs to provide sufficient and comprehensive correction evidence words and use the following symbols to emphasize the evidence words and correction methods:
    - Evidence words must be text segments appearing in the erroneous sentence, surrounded by 【】.
    - Correction methods must be text segments appearing in the corrected sentence, surrounded by {}.

Note: Most examples below contain only one grammatical error, but formal input data usually contains multiple grammatical errors, and you must provide corresponding explanations for each grammatical error. The output must strictly follow the JSON format.

1.Punctuation-level errors: These involve grammatical errors related to punctuation marks.

1.1 Punctuation Redundancy : Refers to inserting punctuation marks unnecessarily. For redundant punctuation errors, first explain the function of the involved punctuation mark, then explain the reason for the redundant punctuation.

Suggested template for explaining redundant punctuation: [Explain the basic usage of redundant punctuation and related evidence words] [Explain the reason for the redundant punctuation] Delete [redundant punctuation] Example of input with redundant punctuation:

```
{
    "error_sentence": "所以一些人说，：“读书一点用处都没有。””,
    "correct_sentence": "所以一些人说：“读书一点用处都没有。””,
    "edit": [
        {
            "src_interval": [6,7],
            "tgt_interval": [6,6],
            "src_content": "，",
            "tgt_content": ""
        }
    ]
}
```

Example of output for redundant punctuation:

```
{
     "edits": [
        {
            "src_interval": [6,7],
            "tgt_interval": [6,6],
            "src_content": "，",
            "tgt_content": "",
            "error_severity": 1,
            "error_type": "标点冗余",
            "error_description": "【：】直接用于【说】等总说性或提示性词语后面，提起下文，没有必要在【：】前插入逗号。应删去【说】之后的冒号。"
        }
    ]
}
```

1.2 Punctuation Missing : Mainly refers to missing punctuation marks that should be present in the sentence, either within or at the end of the sentence. For explaining missing punctuation errors, first identify the evidence words and the missing punctuation mark, then explain the function of the punctuation mark in that context.

Suggested template for explaining missing punctuation: [Explain the basic usage of the missing punctuation and related evidence words] [Explain the reason for the missing punctuation] Add the missing punctuation before/after [evidence words]

Figure 9: The English prompt used for explanation generation.