# EXCGEC: A Benchmark for Edit-Wise Explainable Chinese Grammatical Error Correction

**Jingheng Ye[1*], Shang Qin[1*], Yinghui Li[1], Xuxin Cheng[2], Libo Qin[3], Hai-Tao Zheng[1†], Ying Shen[4], Peng Xing[1], Zishan Xu[1], Guo Cheng[1], Wenhao Jiang[5†]**

[1]Tsinghua University,
[2]Peking University,
[3]Central South University,
[4]Sun Yat-Sen University,
[5]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
{yejh22, qin-s23, liyinghu20}@mails.tsinghua.edu.cn

## Abstract

Existing studies explore the explainability of Grammatical Error Correction (GEC) in a limited scenario, where they ignore the interaction between corrections and explanations and have not established a corresponding comprehensive benchmark. To bridge the gap, this paper first introduces the task of EXplainable GEC (**EXGEC**), which focuses on the integral role of 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 then benchmark several series of LLMs in multi-task learning settings, including post-explaining and pre-explaining. To promote the development of the task, we also build a comprehensive evaluation suite by leveraging existing automatic metrics and conducting human evaluation experiments to demonstrate the human consistency of the automatic metrics for free-text explanations. Our experiments reveal the effectiveness of evaluating free-text explanations using traditional metrics like METEOR and ROUGE, and the inferior performance of multi-task models compared to the pipeline solution, indicating its challenges to establish positive effects in learning both tasks.

**Code & Data** — https://github.com/THUKElab/EXCGEC

## Introduction

Despite the notable advancements in Grammatical Error Correction (GEC) (Bryant et al. 2023; Ye et al. 2023a; Li et al. 2025), 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 (second language)-speakers (Wang et al. 2021). 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 GEC as well as related fields, such as essay scoring (Stahl et al. 2024), intelligent tutoring systems (Montenegro-Rueda et al. 2023).

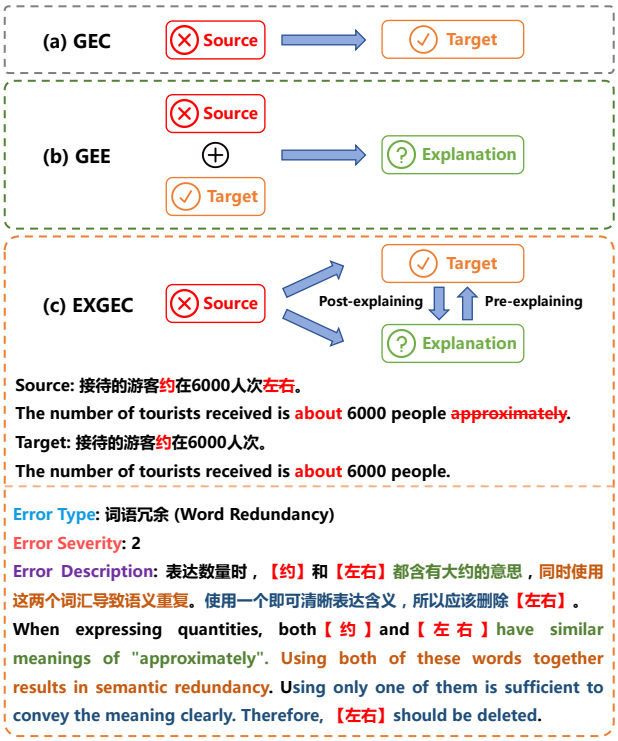


Figure 1: Task definitions of GEC, GEE, and EXGEC. We highlight 【evidence words】, {correction}, linguistic knowledge, error causes, and revision advice parts.

As illustrated in Figure 1, existing tasks like GEC and Grammatical Error Explanation (GEE) typically address either correction or explanation, ignoring the interaction be-

*These authors contributed equally.
†Corresponding Authors.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tween 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 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 edit-wise 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 the recent success of synthetic data generation (Shum, Diao, and Zhang 2023; Whitehouse, Choudhury, and Aji 2023), we employ a semi-automatic dataset construction solution to enhance efficiency, while minimising annotation costs. Initially, we synthesize the evaluation part of EXCGEC by prompting GPT-4 (Liu et al. 2024). Then we hire experienced annotators to filter out invalid data and concurrently provide a detailed analysis of the invalid data, ensuring the quality of our dataset (Ding et al. 2024). We finally obtain 8,216 clean explanation-augmented samples for benchmarking. Additionally, We utilize existing automatic metrics to evaluate the performance. Particularly for error descriptions, we conduct a human evaluation experiment to ascertain the correlation between the metrics and human judgements, thus demonstrating their effectiveness.

Based on the benchmark, we develop EXGEC multi-task 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. However, both of them under-perform the pipeline solution. Moreover, COTE significantly enhances performance by alleviating the alignment workload for the LLMs. Our contributions in this paper are listed as follows:

- We introduce the EXGEC task and establish a corresponding benchmark consisting of a Chinese EXGEC dataset and a comprehensive set of metrics, contributing to the stable development of the field of EXGEC.
- We develop EXGEC baseline models and investigate the abilities of various LLMs using our proposed benchmark.
- 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.

## Related Work

**Explainable GEC.** Exploration of explainable GEC has witnessed a paradigm shift from fine-tuning to prompting (Zhao et al. 2024). EXPECT (Fei et al. 2023) is an explainable GEC dataset annotated with evidence words and error types based on the standard GEC benchmark (Bryant et al. 2019). 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 trained on synthesized data. On the other hand, a similar task called feedback comment generation, focuses on sentence-level explanations. However, it suffers from expensive costs associated with data annotation (Nagata, Inui, and Ishikawa 2020). Furthermore, it is explored with limited access to only a subset of English grammatical error types due to the complexity of the task (Nagata 2019). In conclusion, all these studies do not establish a comprehensive benchmark integrating both the tasks of GEC and GEE, and thus lack in-depth exploration in multi-task learning the both tasks. However, our work is the first to propose a systematic framework for EXCGEC.

**Chinese GEC.** The research on CGEC (Ye et al. 2023a; Ye, Li, and Zheng 2023) has also come a long way recently, along with a series of CGEC datasets (Zhao et al. 2018). Similar to those in English, Chinese grammatical errors can also be categorized into different error types. CLG (Ma et al. 2022) divides Chinese grammatical errors into 6 categories: Structural Confusion, Improper Logicality, Missing Component, Redundant Component, Improper Collocation, and Improper Word Order. However, the taxonomy of CLG is targeted toward grammatical errors made by native speakers and thereby can not cover those made by L2 speakers. To fill the gap, we design a two-tier hierarchical taxonomy, which is capable of covering most grammatical errors.

## Task Definition

### Grammatical Error Correction

GEC (Schneider and McCoy 1998) 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 follows:

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

## Grammatical Error Explanation

GEE (Song et al. 2023) has received much attention recently and has been explored in several methodologies, including sentence-level explanation and edit-wise explanation. Since sentence-level explanations 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, Felice, and Briscoe 2017) and CLEME (Ye et al. 2023b, 2024). 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 $X$ and $Y$. 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:

$$E = f(X, Y), \tag{2}$$

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

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 explanations. However, an extra GEC model is compulsory for GEE models to work, thus resulting in an issue of low efficiency.

## Explainable Grammatical Error Correction

To get rid of the drawbacks brought by the nature of GEE, we propose the EXGEC task which aims to perform both correction and explanation tasks simultaneously. The motivation for combining these two tasks majorly falls on two aspects. First, a branch of existing studies (Wiegreffe and Marasovic 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 the 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 grammatical 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:

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

The inference of post-explaining models is as follows:

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

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

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

**Pre-explaining.** This type of model is trained in converse order, 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 follows:

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

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

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

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

# 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 with the process of data curation, which consists of Explanation Design, Explanation Synthesizing, Explanation Refinement, and Analysis. Then we gain an in-depth understanding of GPT-4 (Achiam et al. 2023) by further analyzing the generated explanations, where we summarize common failure modes in invalid instances. Finally, we explain the evaluation for both the correction and the explanation tasks.

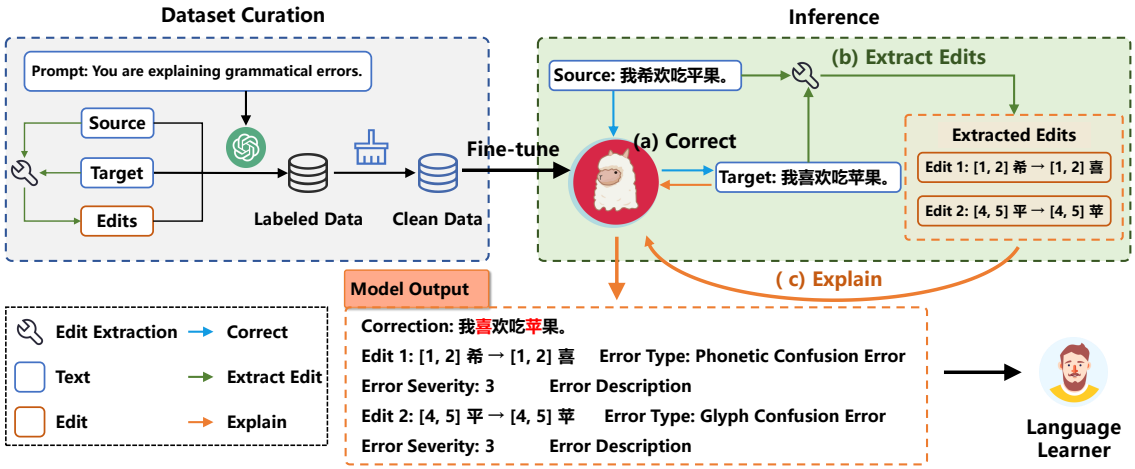


Figure 2: Overview of the benchmark and the model. We show the inference process of a post-explaining model in particular.

| 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 (Phonetic Confusion Error)<br>字形混淆错误 (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 (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.

## 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 descriptions. 1) **Error type labels** allow language learners to comprehend and inductively infer syntax and grammar rules. In particular, we employ a two-tier hierarchical taxonomy including 5 major types and 16 minor types shown in Table 1, inspired by authoritative linguistic books (Huang and Liao 2011; Shao 2016). Detailed descriptions of various error types are included in the supplementary materials. 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 descriptions should meet three nonoverlapping principles: fluency, reasonability (making sense to humans), and faithfulness (targeted to a specific edit). To ensure 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.

## Explanation Synthesizing

Annotating high-quality explanations on 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, including FCGEC (Xu et al. 2022), YACLC (Wang et al. 2021), MuCGEC (Zhang et al. 2022), NaCGEC (Ma et al. 2022), NLPCC (Zhao et al. 2018) and HSK (Zhang 2009). The details are listed in Table 2. We pick out only the samples with *changed* reference sentences to maximize training efficiency (Zhang et al. 2022). We select the reference sentence with the most edits as the target sentence if a sample is annotated with multiple reference sentences. Then, we prompt GPT-4 to generate edit-wise explanations following in-context learning. To ensure the faithfulness of the synthesized explanation, we first extract edits using the toolkit CLEME (Ye et al. 2023b). 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 the

| 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-train** | 5,966 | 2.0 | 38.7 |
| **EXCGEC-dev** | 750 | 2.0 | 38.9 |
| **EXCGEC-test** | 1,500 | 2.0 | 39.2 |
| **EXCGEC** (all) | 8,216 | 2.0 | 38.8 |

Table 2: Dataset statistics of the EXCGEC benchmark.

supplementary materials.

## Explanation Refinement and Analysis

Benefiting from the extensive knowledge acquired during the large-scale pre-training process, GPT-4 can 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 hired 12 native speakers, all of whom are Chinese post-graduated students specializing in Chinese linguistics, to screen out invalid explanations. Before formal annotation, we compile the annotation guidelines and all the annotators receive intensive training. Two authors of the paper, who are also in charge of compiling the annotation guidelines, have made sure that their annotation accuracies are over 90% on testing samples. We make sure that each formal sample is checked by at least two annotators. We finally obtained 8,216 clean samples out of 10,000 samples. We further investigate the failure modes of these invalid explanations, which are provided in the supplementary materials.

## Automatic Metrics

To promote the efficient development of EXGEC systems, we introduce a comprehensive suite of automatic metrics for both correction and explanation parts. Additionally, we conduct a human evaluation experiment in Section Analysis to demonstrate the alignment of the metrics used for assessing error descriptions with human judgments.

**Correction.** We employ CLEME (Ye et al. 2023b) and ChERRANT (Zhang et al. 2022) to evaluate the correction performance. Both are edit-based metrics that output P/R/$F_{0.5}$ scores, which have been proven reliable metrics for GEC on CoNLL-2014 (Ye et al. 2023b).

**Explanation.** Since an edit-wise explanation consists of three critical elements, we define respectively automatic metrics for them. 1) Accuracy 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 the mean absolute error (MAE) to show the deviation of hypothesis error severity levels towards ground truth ones. 3) We employ various metrics for evaluating the free-text explanation descriptions considering both the reproductivity and efficiency, including BLEU (Papineni et al. 2002; Clinciu, Eshghi, and Hastie 2021), METEOR (Banerjee and Lavie 2005), ROUGE (Lin 2004).

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** $\hat{Y}, \hat{E}'$

# Method

**Training.** To streamline the training process covering all the tasks mentioned previously, 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 elements of edit-wise explanations, and provides a consistent and controllable view to distinguish tasks, enabling the model to understand essential task elements and their relations. With this uniform format stipulation, we can train all models using the same smooth cross-entropy loss, regardless of the specific task.

**Inference.** For post-explaining EXGEC models, we design a specific **Co**rrect-**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 in LLMs. Then, we apply CLEME to extract edits. Notably, we merge adjacent edits with a distance of 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.

# Experiments

## Experimental Settings

**Backbones.** We benchmark mainstream LLMs including Qwen-1.5 (Bai et al. 2023), Llama-3 (Touvron et al. 2023), and DeepSeek (Bi et al. 2024). For these LLMs, we experiment with their base and chat (or instruct) versions to inves-

| | 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 | **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 |
| Pre | **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-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-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-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.

tigate whether further alignment training benefits the task. All experimental results are averaged over three runs with different random seeds on EXCGEC-test in Table 2. More training details are reported in the supplementary materials.

**Evaluation.** We obtain the metric results using public toolkits including *ROUGE* (Lin 2004), *NLTK* (Bird and Loper 2004), and *scikit-learn* (Pedregosa et al. 2011). Particularly, we observe many hypothesis edits are not covered by the corresponding reference edits, making it impossible to subsequently evaluate the 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 the hit edits are used to calculate the evaluation outcomes for explanations.

## Results of Multi-task Models

Table 3 presents the main results of multi-task models.

**Post-explaining models outperform pre-explaining models.** Concerning the correction aspect, all post-explaining models consistently 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 complexity for LLMs to directly explain grammatical errors without auxiliary information like target sentences or extracted edits. And once pre-explaining models generate flawed explanations, the ensuing distraction impedes their ability to accurately correct the source text.

**Chat models are superior to 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.

## Ground Truth Results

To examine the isolated performance of multi-task models, we introduce partial ground truth information in advance during the formal inference stage. This is achieved by pre-inserting ground truth corrections or explanations into the decoding phase prior to formal inference. Specifically, we utilize ground truth target texts for post-explaining and evaluate the performance of the explanation task. Conversely, we provide ground truth explanations for pre-explaining and assess the performance of the correction task. This approach enables a detailed analysis of each task's performance under oracle conditions. The results, as depicted in Table 4, reveal that the incorporation of ground truth information significantly enhances performance. Notably, post-explanatory models equipped with ground truth corrections exhibit a marked improvement in explanatory performance across all LLMs. This observation extends to post-explanatory models with ground truth explanations, suggesting that previously generated low-quality content adversely affects subsequent generative processes.

| 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 |
| **GEC-GEE Pipeline** | **32.45 / 23.93 / 30.29** | **40.50 / 19.58 / 33.37** | **72.00** | **63.10** | **65.76** | **32.77** | **0.70** | **16.41** | **40.04** | **51.07 / 24.92 / 35.89** |

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

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

Table 6: Ablation results of COTE from the same 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 7: Correlations between human judgements ($A_1$ and $A_2$) and metrics results for error descriptions.

### Comparison with Pipeline

We compare multi-task models and a GEC-GEE pipeline with COTE in Table 5. It indicates that the pipeline solution can improve both the correction and the explanation performance compared to multi-task models, highlighting the challenges of learning multi-task models for EXCGEC. However, adopting the pipeline solution requires heavy deployment and training costs. We speculate that LLMs with only 7B parameters cannot establish intimate interaction of correction and explanation tasks.

## Analysis

### 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 and provide a case study for different LLMs in the supplementary materials.

**Effect of COTE.** We introduce COTE that provides gold alignment for post-explaining models, thus unburdening LLMs during the inference stage. The impact of COTE is quantitatively 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 6 reveal a huge performance drop if we do not leverage COTE, especially the hit rate and the miss rate. This demonstrates the effectiveness of COTE.

### Human Evaluation for Error Descriptions

We adopt traditional metrics for assessing the quality of generated error descriptions mainly for their reproductivity and efficiency (Clinciu, Eshghi, and Hastie 2021). However, their reliability requires further validation. Therefore, this section attempts to demonstrate the suitability of these metrics through their corrections with human judgments. We assign two human annotators to score the error descriptions generated by all 6 post-explaining models, with the scoring scale from 0 to 100. For each sample, the annotators are instructed to concurrently evaluate all the error descriptions, referencing a gold explanation generated by GPT-4 to guarantee a rigorous and reliable assessment. Additional details are delineated in the supplementary materials.

We report Pearson and Spearson correlations between the metric results and the human judgments in Table 7. We observe the inter-annotator correlations are close to 1, meaning it is relatively easy to determine the quality of error descriptions for human annotators. Most metrics achieve moderate or high correlations with human judgments, which means that it is relatively reasonable to use simple n-grams-based metrics to evaluate the quality of error descriptions efficiently. Among various metrics, ROUGE-1 achieves 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 detailed annotation guidance and rating rules in the supplementary materials.

## Conclusion

We propose and formulate the task of EXGEC, establishing the interaction of correction and explanation tasks. To develop the task, we propose the EXCGEC benchmark, based on which we build baseline models. Extensive experiments and analyses reveal several challenges of the task. We hope this paper can serve as a starting point for future exploration.


## Acknowledgements

This research is supported by National Natural Science Foundation of China (Grant No.62276154), Research Center for Computer Network (Shenzhen) Ministry of Education, the Natural Science Foundation of Guangdong Province (Grant No.2023A1515012914 and 44030024103310080177O), Basic Research Fund of Shenzhen City (Grant No.JCYJ20210324120012033 and GJHZ20240218113603006), the Major Key Project of PCL for Experiments and Applications (PCL2021A06).


## References


Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.;

Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Goldstein, J.; Lavie, A.; Lin, C.-Y.; and Voss, C., eds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics.

Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; Du, Q.; Fu, Z.; et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Bird, S.; and Loper, E. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 214–217. Barcelona, Spain: Association for Computational Linguistics.

Bryant, C.; Felice, M.; Andersen, Ø. E.; and Briscoe, T. 2019. The BEA-2019 Shared Task on Grammatical Error Correction. In Yannakoudakis, H.; Kochmar, E.; Leacock, C.; Madnani, N.; Pilán, I.; and Zesch, T., eds., *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 52–75. Florence, Italy: Association for Computational Linguistics.

Bryant, C.; Felice, M.; and Briscoe, T. 2017. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In Barzilay, R.; and Kan, M.-Y., eds., *Proceedings of ACL*, 793–805. Vancouver, Canada: Association for Computational Linguistics.

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.

Camburu, O.-M.; Rocktäschel, T.; Lukasiewicz, T.; and Blunsom, P. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Clinciu, M.-A.; Eshghi, A.; and Hastie, H. 2021. A Study of Automatic Metrics for the Evaluation of Natural Language Explanations. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of EACL*, 2376–2387. Online: Association for Computational Linguistics.

Ding, B.; Qin, C.; Zhao, R.; Luo, T.; Li, X.; Chen, G.; Xia, W.; Hu, J.; Luu, A. T.; and Joty, S. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. *arXiv preprint arXiv:2403.02990*.

Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; et al. 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9): 1–33.

Fei, Y.; Cui, L.; Yang, S.; Lam, W.; Lan, Z.; and Shi, S. 2023. Enhancing Grammatical Error Correction Systems with Explanations. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of ACL*, 7489–7501. Toronto, Canada: Association for Computational Linguistics.

Gao, C.; Zhang, W.; Chen, G.; and Lam, W. 2023. JsonTuning: Towards Generalizable, Robust, and Controllable Instruction Tuning. *arXiv preprint arXiv:2310.02953*.

Hartmann, M.; and Sonntag, D. 2022. A survey on improving NLP models with human explanations. In Andreas, J.; Narasimhan, K.; and Nematzadeh, A., eds., *Proceedings of the First Workshop on Learning with Natural Language Supervision*, 40–47. Dublin, Ireland: Association for Computational Linguistics.

He, X.; Wu, Y.; Camburu, O.-M.; Minervini, P.; and Stenetorp, P. 2023. Using Natural Language Explanations to Improve Robustness of In-context Learning for Natural Language Inference. *arXiv preprint arXiv:2311.07556*.

Huang, B.; and Liao, X. 2011. *Modern Chinese (Updated Fifth Edition)*. Higher Education Press, Beijing, China.

Kaneko, M.; Takase, S.; Niwa, A.; and Okazaki, N. 2022. Interpretability for Language Learners Using Example-Based Grammatical Error Correction. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of ACL*, 7176–7187. Dublin, Ireland: Association for Computational Linguistics.

Li, D.; Hu, B.; Chen, Q.; Xu, T.; Tao, J.; and Zhang, Y. 2022. Unifying model explainability and robustness for joint text classification and rationale extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10947–10955.

Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P. S.; and He, L. 2020. A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*.

Li, S.; Chen, J.; yelong shen; Chen, Z.; Zhang, X.; Li, Z.; Wang, H.; Qian, J.; Peng, B.; Mao, Y.; Chen, W.; and Yan, X. 2024. Explanations from Large Language Models Make Small Reasoners Better. In *2nd Workshop on Sustainable AI*.

Li, Y.; Ma, S.; Chen, S.; Huang, H.; Huang, S.; Li, Y.; Zheng, H.-T.; and Shen, Y. 2025. Correct like humans: Progressive learning framework for Chinese text error correction. *Expert Systems with Applications*, 265: 126039.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, R.; Wei, J.; Liu, F.; Si, C.; Zhang, Y.; Rao, J.; Zheng, S.; Peng, D.; Yang, D.; Zhou, D.; et al. 2024. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*.

Ma, S.; Li, Y.; Sun, R.; Zhou, Q.; Huang, S.; Zhang, D.; Yangning, L.; Liu, R.; Li, Z.; Cao, Y.; Zheng, H.; and Shen, Y. 2022. Linguistic Rules-Based Corpus Generation for Native Chinese Grammatical Error Correction. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 576–589. Association for Computational Linguistics.

Montenegro-Rueda, M.; Fernández-Cerero, J.; Fernández-Batanero, J. M.; and López-Meneses, E. 2023. Impact of the implementation of ChatGPT in education: A systematic review. *Computers*, 12(8): 153.

Nagata, R. 2019. Toward a task of feedback comment generation for writing learning. In *Proceedings of EMNLP-IJCNLP*, 3206–3215.

Nagata, R.; Hagiwara, M.; Hanawa, K.; Mita, M.; Chernodub, A.; and Nahorna, O. 2021. Shared task on feedback comment generation for language learners. In *Proceedings of the 14th International Conference on Natural Language Generation*, 320–324.

Nagata, R.; Inui, K.; and Ishikawa, S. 2020. Creating corpora for research in feedback comment generation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 340–345.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

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

Schneider, D.; and McCoy, K. F. 1998. Recognizing Syntactic Errors in the Writing of Second Language Learners. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, 1198–1204. Montreal, Quebec, Canada: Association for Computational Linguistics.

Shao, J. 2016. *General Theory of Modern Chinese*. Shanghai Educational Publishing House, Shanghai, China.

Shum, K.; Diao, S.; and Zhang, T. 2023. Automatic Prompt Augmentation and Selection with Chain-of-Thought from Labeled Data. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12113–12139. Singapore: Association for Computational Linguistics.

Song, Y.; Krishna, K.; Bhatt, R.; Gimpel, K.; and Iyyer, M. 2023. Gee! grammar error explanation with large language models. *arXiv preprint arXiv:2311.09517*.

Stahl, M.; Biermann, L.; Nehring, A.; and Wachsmuth, H. 2024. Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. *arXiv preprint arXiv:2404.15845*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, Y.; Kong, C.; Yang, L.; Wang, Y.; Lu, X.; Hu, R.; He, S.; Liu, Z.; Chen, Y.; Yang, E.; et al. 2021. YACLC: a Chinese learner corpus with multidimensional annotation. *arXiv preprint arXiv:2112.15043*.

Wang, Y.; Zhong, W.; Li, L.; Mi, F.; Zeng, X.; Huang, W.; Shang, L.; Jiang, X.; and Liu, Q. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Whitehouse, C.; Choudhury, M.; and Aji, A. F. 2023. LLM-powered Data Augmentation for Enhanced Cross-lingual Performance. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of EMNLP*, 671–686. Singapore: Association for Computational Linguistics.

Wiegreffe, S.; and Marasovic, A. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Xu, L.; Wu, J.; Peng, J.; Fu, J.; and Cai, M. 2022. FCGEC: Fine-Grained Corpus for Chinese Grammatical Error Correction. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 1900–1918. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Ye, J.; Li, Y.; Li, Y.; and Zheng, H. 2023a. MixEdit: Revisiting Data Augmentation and Beyond for Grammatical Error Correction. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, 10161–10175. Association for Computational Linguistics.

Ye, J.; Li, Y.; and Zheng, H. 2023. System Report for CCL23-Eval Task 7: THU KELab (sz) - Exploring Data Augmentation and Denoising for Chinese Grammatical Error Correction. In Sun, M.; Qin, B.; Qiu, X.; Jiang, J.; and Han, X., eds., *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, 262–270. Harbin, China: Chinese Information Processing Society of China.

Ye, J.; Li, Y.; Zhou, Q.; Li, Y.; Ma, S.; Zheng, H.-T.; and Shen, Y. 2023b. CLEME: Debiasing Multi-reference Evaluation for Grammatical Error Correction. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of EMNLP*, 6174–6189. Association for Computational Linguistics.

Ye, J.; Xu, Z.; Li, Y.; Cheng, X.; Song, L.; Zhou, Q.; Zheng, H.-T.; Shen, Y.; and Su, X. 2024. CLEME2. 0: Towards More Interpretable Evaluation by Disentangling Edits for Grammatical Error Correction. *arXiv preprint arXiv:2407.00934*.

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

Zhang, Y.; Li, Z.; Bao, Z.; Li, J.; Zhang, B.; Li, C.; Huang, F.; and Zhang, M. 2022. MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction. *arXiv preprint arXiv:2204.10994*.

Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; and Du, M. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2): 1–38.

Zhao, Y.; Jiang, N.; Sun, W.; and Wan, X. 2018. Overview of the nlpcc 2018 shared task: Grammatical error correction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 439–445.