Towards Explainable Chinese Native Learner Essay Fluency Assessment: Dataset, Tasks, and Method

Xinshu Shen[1], **Hongyi Wu**[1], **Yadong Zhang**[1], **Man Lan**[1,2,3*], **Xiaopeng Bai**[2,3],
Shaoguang Mao[4], **Yuanbin Wu**[1,2,3], **Xinlin Zhuang**[1], **Li Cai**[1]
[1]School of Computer Science and Technology, East China Normal University
[2]Shanghai Institute of Artificial Intelligence for Education, East China Normal University
[3]Department of Chinese Language and Literature, East China Normal University
[4]Microsoft Research Asia
{xinshushen, hongyiwu}@stu.ecnu.edu.cn
{mlan, ybwu}@cs.ecnu.edu.cn, xpbai@zhwx.ecnu.edu.cn
shaoguang.mao@microsoft.com

Abstract

Grammatical Error Correction (GEC) is a crucial technique in Automated Essay Scoring (AES) for evaluating the fluency of essays. However, in Chinese, existing GEC datasets often fail to consider the importance of specific grammatical error types within compositional scenarios, lack research on data collected from native Chinese speakers, and largely overlook cross-sentence grammatical errors. Furthermore, the measurement of the overall fluency of an essay is often overlooked. To address these issues, we present **CEFA** (Chinese Essay Fluency Assessment), an extensive corpus that is derived from essays authored by native Chinese-speaking primary and secondary students and encapsulates essay fluency scores along with both coarse and fine-grained grammatical error types and corrections. Experiments employing various benchmark models on CEFA substantiate the challenge of our dataset. Our findings further highlight the significance of fine-grained annotations in fluency assessment and the mutually beneficial relationship between error types and corrections[1].

1 Introduction

Essay fluency refers to the coherence of a sentence or a whole composition, as well as grammatical accuracy (Yang et al., 2012), serving as a foundational component in Automated Essay Scoring (AES). The study of essay fluency has significant applications in fields such as education (Gong et al., 2021), text generation (Ahn et al., 2016) and publishing (Wang et al., 2021).

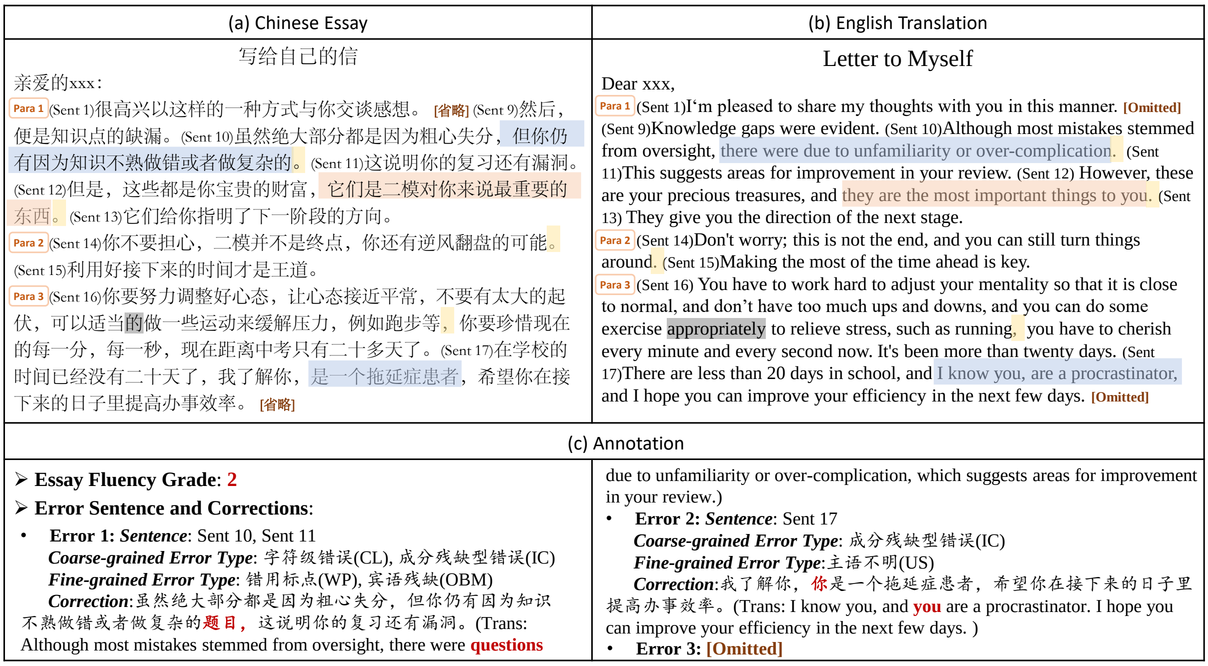
Recent advancements in AES have integrated Grammatical Error Correction (GEC) to improve explainability (Tsai et al., 2020; Gong et al., 2021), with GEC focusing on automatic text error correction (Bryant et al., 2022). In Chinese AES, the prevalent Chinese GEC (CGEC) categorizes errors into four modification types (Gong et al., 2021) and make corrections. Subsequently, an overall score of the essay is conducted based on the errors and other linguistic features. This method, while adding some explainability to the scoring process, offers limited insights for students seeking to understand complex grammatical rules. Additionally, it lacks a distinct fluency score to assess the specific impact of grammatical errors on essay fluency and the overall level of fluency in the essay, which is a crucial component in essay grading.

The existing CGEC dataset is not directly applicable for assessing essay fluency. **Primarily**, most CGEC methods rely on corpora from Chinese-as-a-second-language (CSL) learners, who are more prone to lexical confusion errors, such as confusing "关爱" (care and love) and "爱情" (romantic love), both translated as "love" in English (Wang et al., 2022), which is seldom seen among native speakers. **Additionally**, existing corpora often derive from online texts, which typically do not adhere to language usage norms and grammars. **Moreover**, the definition of error types is not sufficiently detailed. Recent datasets either predominantly focus on orthographic errors like typos (Zhang et al., 2022a, 2023), or solely target syntactic errors like constituent omissions (Xu et al., 2022), which lacks comprehensiveness and diversity. **Lastly**, existing datasets lack annotations for cross-sentence errors (Chollampatt et al., 2019; Yuan and Bryant, 2021), which are common in documents, as illustrated in Figure 1(c) Error 1.

To tackle the issues, we propose an detailed assessment guideline for automatic essay assessment in fluency and developed the **C**hinese **E**ssay

[*] Corresponding author.
[1] Our code and dataset are publicly available at https://github.com/cubenlp/CEFA

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 15515–15528
November 12-16, 2024 ©2024 Association for Computational Linguistics



(a) Chinese Essay

写给自己的信

亲爱的xxx：

Para 1 (Sent 1)很高兴以这样的一种方式与你交谈感想。 [省略] (Sent 9)然后，便是知识点的缺漏。(Sent 10)虽然绝大部分都是因为粗心失分，但你仍有因为知识不熟做错或者做复杂的。(Sent 11)这说明你的复习还有漏洞。(Sent 12)但是，这些都是你宝贵的财富，它们是二模对你来说最重要的东西。(Sent 13)它们给你指明了下一阶段的方向。

Para 2 (Sent 14)你不要担心，二模并不是终点，你还有逆风翻盘的可能。(Sent 15)利用好接下来的时间才是王道。

Para 3 (Sent 16)你要努力调整好心态，让心态接近平常，不要有太大的起伏，可以适当的做一些运动来缓解压力，例如跑步等，你要珍惜现在的每一分，每一秒，现在距离中考只有二十多天了。(Sent 17)在学校的时间已经没有二十天了，我了解你，是一个拖延症患者，希望你在接下来的日子里提高办事效率。 [省略]

(b) English Translation

Letter to Myself

Dear xxx,

Para 1 (Sent 1)I'm pleased to share my thoughts with you in this manner. [Omitted] (Sent 9)Knowledge gaps were evident. (Sent 10)Although most mistakes stemmed from oversight, there were due to unfamiliarity or over-complication. (Sent 11)This suggests areas for improvement in your review. (Sent 12) However, these are your precious treasures, and they are the most important things to you. (Sent 13) They give you the direction of the next stage.

Para 2 (Sent 14)Don't worry; this is not the end, and you can still turn things around. (Sent 15)Making the most of the time ahead is key.

Para 3 (Sent 16) You have to work hard to adjust your mentality so that it is close to normal, and don't have too much ups and downs, and you can do some exercise appropriately to relieve stress, such as running, you have to cherish every minute and every second now. It's been more than twenty days. (Sent 17)There are less than 20 days in school, and I know you, are a procrastinator, and I hope you can improve your efficiency in the next few days. [Omitted]

(c) Annotation

➢ **Essay Fluency Grade**: 2

➢ **Error Sentence and Corrections**:

- **Error 1: *Sentence***: Sent 10, Sent 11
  Coarse-grained Error Type: 字符级错误(CL), 成分残缺型错误(IC)
  Fine-grained Error Type: 错用标点(WP), 宾语残缺(OBM)
  Correction:虽然绝大部分都是因为粗心失分，但你仍有因为知识不熟做错或者做复杂的题目，这说明你的复习还有漏洞。(Trans: Although most mistakes stemmed from oversight, there were questions due to unfamiliarity or over-complication, which suggests areas for improvement in your review.)
- **Error 2: *Sentence***: Sent 17
  Coarse-grained Error Type: 成分残缺型错误(IC)
  Fine-grained Error Type:主语不明(US)
  Correction:我了解你，你是一个拖延症患者，希望你在接下来的日子里提高办事效率。(Trans: I know you, and you are a procrastinator. I hope you can improve your efficiency in the next few days.)
- **Error 3: [Omitted]**

Figure 1: Example of CEFA annotation: In (a) and (b), highlighted sections mark errors. Colors distinguish error types: blue for incomplete component error (IC), yellow for character-level errors (CL), and orange for incorrect constituent combination error (ICC). (c) offers detailed annotations, with red in "***Correction***" indicating changes.

Fluency **A**ssessment (**CEFA**) corpus. This dataset addresses limitations in prior work: **Firstly**, it simultaneously annotates essay fluency grades, grammatical error types and the corrected sentences, which facilitates a comprehensive and detailed evaluation of the essay in fluency. **Secondly**, it encompasses 5 coarse-grained and 18 fine-grained grammatical error types, providing a basis for scoring and correction, and offering teachers and students precise insights into writing issues and targeted feedback. **Finally**, it originates from essays written by native primary and secondary school students, encompassing a diverse range of topics, genres, and score ranges, and annotates errors from document-level perspectives, which is especially beneficial for a more in-depth study of CGEC.

To assess the complexity of our **CEFA** dataset, we explored several baseline models and large language models (LLMs) on our dataset. The results show that our dataset is challenging. Furthermore, we investigated the impact of detailed annotation on fluency grading, as well as the mutually benefits between grammatical error types and corrections through experiments. The findings emphasize the importance of fine-grained annotations and the strong mutual benefit between error types and corrections.

We summarize our contributions as follows:

- We develop a pioneering dataset CEFA for automatic essay fluency assessment, including fine-grained annotations for various aspects related to essay fluency based on native students' essays. Not only offers valuable data resources for CGEC but facilitates in-depth essay assessments.

- We provide comprehensive benchmarks for each task, investigating the performance of current methods and providing a reference point for future research.

- Through experiments, we explore the value of detailed annotations for grading, the optimal benefit between error types and corrections, and the significance of cross-sentence errors.

2 Related Work

2.1 Automated Essay Scoring

Automated essay scoring (AES) is a computer-based assessment system that automatically scores or grades essays by considering appropriate features (Ramesh and Sanampudi, 2022). Originally designed to assign grades or scores to essays, AES now assesses essays to reduce teachers' grading workload, enhance students' writing skills, including offering personalized feedback by evaluating aspects like advancing expressions, grammatical accuracy, and tailored comments (Wu et al., 2023; Zhuang et al., 2024), driven by the expansion of online education and rising manual grading costs (Zhang et al., 2022b).

2.2 Automatic Essay Fluency Assessment

Essay fluency is an important feature of automated essay assessment, which refers to the measure of the normative use of grammar and the coherence of the essay. The assessment of it was commonly treated as a singular natural language processing (NLP) task. These methods integrate linguistic features like sentence length and vocabulary complexity to provide scores or grades for fluency (Mim et al., 2021; Yang et al., 2019), or use language models to calculate sentence probabilities for coherence evaluation (Kann et al., 2018). E-rater (Attali and Burstein, 2004) provides grammar errors as an aid in scoring, but neglects corrections for improvement. Some also treated as GEC task, correcting spelling and grammar errors (Gong et al., 2021; Tsai et al., 2020). Specifically, they correct errors from four perspectives: insertion, modification, deletion and reordering. However, this error definition fails to measure errors from the abstract grammar aspect, leaving both students and teachers unable to clearly grasp the issues in writing.

2.3 Grammatical Error Correction

The GEC task aims to automatically detect and correct grammatical errors in sentences. Despite numerous datasets and methods for English GEC, CGEC resources are limited, with only four publicly accessible datasets: CTC-Qua (Zhao et al., 2022), CCTC (Wang et al., 2022), FCGEC (Xu et al., 2022) and NaSGEC (Zhang et al., 2023).

Unlike online texts, written texts place more emphasis on linguistic norms and conventions of language usage, making the study of grammatical errors in written context more rigorous and precise. However, only a subset of FCGEC and NaSGEC is sourced from writing text in educational field. FCGEC consists of multi-choice questions from public school Chinese examinations. It defines 7 error types for annotation, but neglects simple grammatical errors such as typos and punctuation mistakes, making the error categorization system not comprehensive. NaSGEC is a multi-domain CGEC dataset, derived from native texts, with data sourced from online texts and sentence error determination questions in Chinese language exams. While it often constructed for the purpose of practicing specific grammar knowledge and may differ from real writing scenarios.

Coarse-grained Types	Fine-grained Types
Character-Level Error (CL)	Word Missing (WM), Typographical Error (TE), Missing Punctuation (MP), Wrong Punctuation (WP)
Redundant Component Error (RC)	Subject Redundancy (SR), Particle Redundancy (PR), Statement Repetition(SRP), Other Redundancy (OR)
Incomplete Component Error (IC)	Unknown Subject (US), Predicate Missing (PM), Object Missing (OBM), Other Missing (OTM)
Incorrect Constituent Combination Error (ICC)	Inappropriate Subject-Verb Collocation (ISVC), Inappropriate Verb-Object Collocation (IVOC), Inappropriate Word Order (IWO), Inappropriate Other Collocation (IOC)
Illogical (IL)	Linguistic Illogicality (LIL), Factual Illogicality (FIL)

Table 1: Our guideline adopts 5 coarse-grained and 18 fine-grained error types.

Set	Essay	Error Sent	Chars/Sent	Edits/Ref	Multi Label	Cross Sent
All	501	4,258	46.18	2.80	37.88%	782
Train	350	2,981	45.88	2.74	38.27%	553
Dev	76	630	47.39	2.74	39.31%	106
Test	75	647	46.40	2.93	35.69%	123

Table 2: Data statistics of CEFA. **Chars/Sent** indicates the average number of characters per sentence, **Edits/Ref** represents the average edit distance per sentence compared to the original sentence, **Multi Label** signifies the proportion of sentences with multiple labels among those containing errors, and **Cross Sent** indicates the number of cross-sentence errors.

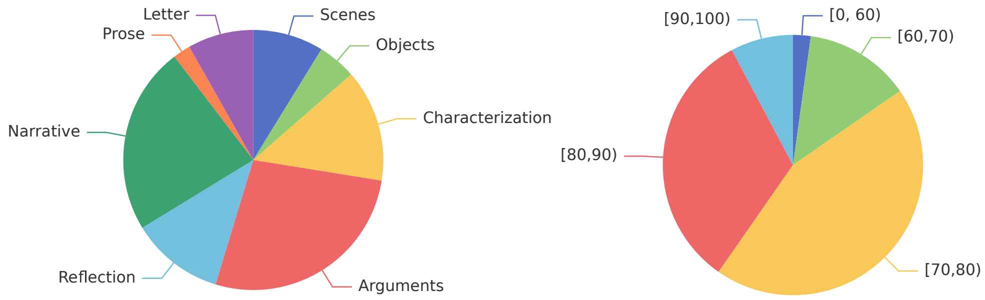
3 Dataset Construction

3.1 Data Collection

The dataset was derived from essays written by students from three local primary and secondary school. We collected around 700 essays from exams and daily practices, covering various writing topics and written by students of different writing levels. 501 essays were screened for further manual annotation, ensuring a diverse representation in terms of grades, genre, and overall scores assigned by Chinese teachers. The distribution of essay genres is shown in Figure 2a, covering eight genres. Figure 2b illustrates the distribution of score ranges, which represent the overall marks assigned to each essay by teachers.

3.2 Annotation Format

In our corpus, each essay consists of a title and body. For each essay, our annotation comprises three components: grading fluency score, identifying error types, and correcting.



(a) Essay genre distribution.

(b) Essay score distribution.

Figure 2: (a) displays the distribution of the 501 essays used to construct the dataset by genre, covering a total of 8 essay genres. (b) shows the distribution of the essays used for annotation in terms of score.

Error Type		Train Num (Perc.)	Dev Num (Perc.)	Test Num (Perc.)
Coarse	Fine			
CL	WM	235(5.15%)	47(4.90%)	31(3.29%)
	TE	1169(25.62%)	251(26.15%)	256(27.21%)
	MP	452(9.91%)	88(9.17%)	78(8.29%)
	WP	1183(25.93%)	250(26.04%)	281(29.86%)
RC	SR	17(0.37%)	4(0.42%)	4(0.43%)
	PR	122(2.67%)	19(1.98%)	22(2.34%)
	SRP	21(0.46%)	4(0.42%)	3(0.32%)
	OR	476(10.43%)	98(10.21%)	75(7.97%)
IC	US	316(6.93%)	76(7.92%)	81(8.61%)
	PM	43(0.94%)	11(1.15%)	10(1.06%)
	OBM	65(1.42%)	14(1.46%)	14(1.49%)
	OTM	127(2.78%)	24(2.50%)	25(2.66%)
ICC	ISVC	3(0.07%)	3(0.31%)	2(0.21%)
	IVOC	47(1.03%)	4(0.42%)	3(0.32%)
	IWO	138(3.02%)	21(2.19%)	19(2.02%)
	IOC	138(3.02%)	40(4.17%)	34(3.61%)
IL	FIL	2(0.04%)	1(0.10%)	2(0.21%)
	LIL	9(0.20%)	5(0.52%)	1(0.11%)

Table 3: Distribution of error types in CEFA. **Train/Dev/Test Num (Perc.)** denotes the count and percentage of each type in train/dev/test set.

3.2.1 Essay Fluency Grading

The fluency of an essay is graded as excellent, average, and unsatisfactory. According to the definition of fluency (Yang et al., 2012), we divided the scoring criteria into two parts: the smoothness of the essay and the standardization of language use, which includes native speakers' language intuition and the types and quantities of grammatical errors. It is worth noting that this pertains to the scoring of the essay's fluency in writing, rather than the overall evaluation of the essay. More details are shown in Appendix A.1.

3.2.2 Error Types

Based on prior annotation standards in CGEC (Zhang et al., 2022a; Xu et al., 2022) and the *National Curriculum Standards for Compulsory Education: Chinese Language*, we devise a more comprehensive grammatical error annotation schema. Specifically, we categorize writing errors into character-level and component-level, further subdividing into 5 coarse and 18 fine-grained types, as shown in Table 1. More detailed definitions and examples are shown in Appendix A.2. Annotators identify and label error sentences based on our schema for fine-grained errors. It's worth noting that one sentence may contain multiple errors, requiring annotators to mark all error types within it. This multifaceted annotation allows for a detailed and comprehensive evaluation of each essay.

3.2.3 Correction

GEC annotation employs two paradigms: error coded and rewriting. The former suffers from inconsistent error span definitions and cumbersome modifications for complex sentences, affecting annotation quality. The later offers greater flexibility, which also may hinder the ability to constrain annotators and achieve smooth, minimal changes (Sakaguchi et al., 2016). Therefore, we merge both methods. For character-level errors, we follow the error coded and annotate the index of the incorrect character and the modified character separately. For component-level errors, we use the rewriting paradigm to deal flexibly with complex revisions and add edit distance as a constraint.

3.3 Annotation Process

The annotation team comprised four undergraduates, four postgraduates majoring in linguistics, and four expert reviewers with Chinese teaching experience. During annotation, we divided the data into five groups, each annotated by both one undergraduate and one graduate student, with subsequent expert review. Notably, the first group of data was annotated by four undergraduate students and four graduate students, and then reviewed by four experts. Additionally, our annotation team possess a

Genre	Fluency Grade (%)		
	Excellent	Average	Unsatisfactory
Scenes	72.73	22.73	4.55
Objects	79.17	12.50	8.33
Characterization	28.57	58.57	12.86
Arguments	36.03	51.47	12.50
Reflection	34.48	51.72	13.79
Narrative	27.35	41.88	30.77
Prose	81.82	18.18	0.00
Letter	51.22	41.46	7.32
Total	40.32	44.31	15.37

Table 4: Distribution of fluency grades across different genres, presented as percentages.

deep understanding of language structure, grammar rules, and linguistic expression.

The complexity of our detailed error types and the allowance for diverse corrections pose challenges for annotation. Therefore, we conduct intensive training sessions for annotators before annotation, and hold multiple discussions during the annotation process to ensure the quality. Overall, the annotation process lasted for three months and resulted in the annotation of 501 essays.

3.4 Data Statistics

Our dataset includes 501 essays with 9,912 original sentences, of which 4,258 contained errors and underwent modification. We used 350 essays as the training set, 76 essays as the validation set, and 75 essays as the test set, and the distribution of data can be found in Table 2. Additionally, Table 3 provides a detailed distribution of coarse and fine-grained error types in the dataset. Furthermore, in Table 4, we provide an illustration of the distribution of essay fluency scores (Excellent, Average, Unsatisfactory) across different essay genres.

3.5 Inner Annotator Agreements

To verify annotation quality, we calculated the Inter-Annotator Agreement using Cohen's Kappa for *Essay Fluency Grading* and *Error Types* tasks, and $F_{0.5}$ for *Correction* task, with scores of 0.61, 0.59, and 62.12% for each task. Details are shown in Appendix B.

3.6 Ethical Issues

We've anonymized the data by removing personal details like names and class information before annotation. All annotators and expert reviewers were paid for their work. Besides, we have obtained the permission of the authors and their guardians for all essays used for annotation and publication. We are sincerely grateful for their support.

4 Experiments

4.1 Tasks

Our task comprises three subtasks: `Essay Fluency Grading` for assessing overall essay fluency, `Error Type Identification` for identifying coarse and fine-grained grammatical errors in sentences, noting their potential multi-label nature due to multiple error types, and `Wrong Sentence Rewriting` for rewriting the incorrect sentences for correction.

4.2 Baseline and Metrics

We use the state-of-the-art (SOTA) pre-trained language models (PLMs) in classification tasks like BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) as benchmark models for grading and error identification task. For wrong sentence rewriting task, we establish baselines with models like Chinese BART (Shao et al., 2021) due to its similarity to the pre-training task and our correction task, and Large Language Models (LLMs) including ChatGLM (Du et al., 2022), Baichuan (Baichuan, 2023) and ChatGPT (OpenAI, 2022), noted for their text generation capabilities. We also evaluated the performance of LLMs in the first two tasks. For ChatGPT, both zero-shot and few-shot learning are used for all tasks. For ChatGLM and Baichuan, we fine-tune it with LoRA (Hu et al., 2021). Details of prompts and configurations are shown in Appendix D.

Essay Fluency Grading: We frame this problem as a classification task and employed PLMs mentioned previously as our baselines. We evaluate model performance using Precision (P), Recall (R), F_1, Accuracy (Acc) and Quadratic weighted Kappa (QWK) (Vanbelle, 2016).

Error Type Identification: We fine-tune various PLMs on our training dataset, leveraging their powerful language modeling capabilities. Furthermore, we explored the performance of other novel models in CGEC on our dataset like FCGEC (Xu et al., 2022). For evaluation, we assess our models from both coarse and fine-grained perspectives, utilizing P, R, Micro F_1 and Macro F_1 as our evaluation metrics.

Wrong Sentence Rewriting: Inspired by GEC task, we compare two mainstream correction models: Seq2Edit and Seq2Seq model, on our dataset.

For Seq2Edit, we use the SOTA model, GECToR (Omelianchuk et al., 2020) and STG-Joint (Xu et al., 2022), as our baselines. For Seq2Seq, we fine-tune Chinese BART on our dataset.

For evaluation, the similarity with the ground truth is matched. On the other hand, given the fact that there can be multiple correct corrections for a given sentence, the corrections generated by models may differ from the gold corrections. To address this, we employ language models (LMs) to measure the fluency of the generated corrections. Furthermore, in order to prevent over correction that would significantly alter the original text, we incorporate the Levenshtein distance measure. By minimizing the alterations, we respect the unique expression of the student writer, while correcting their grammatical mistakes. In a word, we evaluate the results of the model from two perspectives:

Comparison with ground truth. We employ three evaluation metrics: **1)** Exact Match (EM): calculates the percentage of correct sentences generated by the model that exactly match the gold references; **2)** Edit metrics proposed by MuCGEC : converts error-correct sentence pairs into operations, compares the model's output operations with the correct references, and calculates the highest scores for F$_{0.5}$; **3)** BLEU: measures the overlap between the model-generated sentences and the correct references.

Correctness and reasonableness of results. We use three metrics: **1)** Perplexity(PPL): measures the quality of rewritten sentences by BERT (Devlin et al., 2018). **2)** BERTScore (Zhang et al., 2019): measures the similarity between the rewritten sentence and the original sentence. **3)** Levenshtein Distance (LD): calculates the edit distance between the rewritten sentence and the original one.

4.3 Implementation Details

For PLMs, we use BERT-Base-Chinese and Chinese-RoBERTa-wwm (Cui et al., 2020) and adopt AdamW optimizer (Loshchilov and Hutter, 2017) with the learning rate of $2e^{-5}$ to update the model parameters and set batch size as 16. For Baichuan, we use Baichuan2-7B-Base as our baseline model. For ChatGLM, we use ChatGLM2-6B. For LLMs, we fine-tuned individually on each task and employed LoRA with the rank parameter set to 8 and the alpha parameter set to 32.

All our experiments are performed on RTX 3090. All other parameters are initialized with the default values in PyTorch Lightning[2], and our model is all implemented by Transformers[3].

Model	**P(%)**	**R(%)**	**F$_1$(%)**	**Acc(%)**	**QWK**
BERT$_{base}$	56.74	46.97	46.76	52.98	0.3868
RoBERTa$_{base}$	54.97	**58.71**	49.70	49.36	0.3961
BERT$_{large}$	55.25	49.09	49.08	53.64	**0.4027**
RoBERTa$_{large}$	56.31	53.94	**54.58**	57.62	0.3830
ChatGPT$_{0-shot}$	56.53	33.54	27.05	42.38	0.1159
ChatGPT$_{1-shot}$	50.41	38.38	38.09	44.37	0.1650
ChatGLM$_{ft}$	47.62	42.32	40.62	46.61	0.2150
Baichuan$_{ft}$	**59.96**	54.86	54.24	**62.67**	0.2386

Table 5: Results for Essay Fluency Grading task.

4.4 Results and Analysis

4.4.1 Essay Fluency Grading

Table 5 presents the best performances of different models on `Essay Fluency Grading` task. It is worth noting that models demonstrate relatively poor performance on this three-classification task. Firstly, the mediocre IAA score (60.36%) observed in the annotation highlights the inherent difficulty of this task. This is primarily due to the subjective nature of the annotation task, which is also influenced by the quality of essays previously annotated by annotators. Secondly, grading is a complex and diverse task, making it difficult for PLMs to learn subtle distinctions solely through training and fintuning.

In testing ChatGPT, ChatGLM and Baichuan on the task, we found few-shot generally outperformed zero-shot. Additionally, we noted a tendency of LLMs to assign the "Excellent" rating, possibly because they lean towards a gentler teaching style. Furthermore, although the fine-tuned Baichuan exhibits better performance attributed to the powerful language understanding capabilities of LLMs, there is still a significant gap between LLMs and human-annotated results, highlighting the need for further exploration when applying LLMs to tasks that involve subjective factors.

4.4.2 Error Type Identification

Table 6 illustrate the main results on `Error Type Identification` task, in terms of both coarse and fine-grained aspects. Detailed results can be found in the Appendix C. BERT and RoBERTa perform well due to their outstanding language understanding capabilities and suitability for the task.

[2] https://github.com/Lightning-AI/lightning

[3] https://github.com/huggingface/transformers

Model	Coarse-grained		Fine-grained	
	Micro $\mathbf{F}_1$	Macro $\mathbf{F}_1$	Micro $\mathbf{F}_1$	Macro $\mathbf{F}_1$
FCGEC	69.25	29.71	44.90	9.52
BERT	69.58	31.29	54.84	15.14
RoBERTa	**70.34**	**34.75**	**56.16**	**18.63**
ChatGPT$_{0-shot}$	15.41	13.05	10.42	7.27
ChatGPT$_{3-shot}$	25.49	16.96	12.40	8.51
ChatGLM$_{ft}$	67.75	31.35	49.42	15.50
Baichuan$_{ft}$	65.61	30.01	50.88	13.13

Table 6: Comparison of performance on coarse and fine-grained error type identification. The PLMs involved are all based on the base version.

Model	EM	$\mathbf{F}_{0.5}$	BLEU-4	BERTScore	LD	PPL
GECToR	11.47	40.03	90.01	96.95	**0.44**	3.16
STG-Joint	12.84	26.21	88.61	96.94	1.80	3.32
BART	18.08	41.21	90.25	97.84	1.67	3.03
ChatGPT$_{0-shot}$	5.56	16.93	76.74	94.38	8.19	3.79
ChatGPT$_{3-shot}$	4.64	17.72	79.81	95.60	5.64	2.94
ChatGLM$_{ft}$	16.45	40.61	90.50	97.63	1.52	3.12
Baichuan$_{ft}$	**22.10**	**41.91**	**90.99**	**97.95**	1.99	**2.94**

Table 7: Results on the Wrong Sentence Rewriting task.

Similarly, for ChatGPT, the few-shot performance is better than zero-shot. Additionally, the LLMs without fintuning demonstrate inferior performance across two levels of granularity, indicating that our task presents a certain degree of challenge to LLMs. The performance of the fine-tuned LLMs still exhibits some gaps compared to RoBERTa. This can be attributed to the nature of our identification task, where a sentence might have multiple error types, where each category label is independent. When employing generative models for classification, there's a necessity to serialize multiple labels, leading to a scenario where predictions of subsequent labels are influenced by preceding ones. This misalignment deviates from the objectives of the identification task.

4.4.3 Wrong Sentence Rewriting

Table 7 shows the `Wrong Sentence Rewriting` task results. GECToR, using a sequence labeling approach, aims for minimal input changes, yielding lower LD values but possibly resulting in less fluent sentences, as indicated by higher PPL scores. STG-Joint designs 3 modules to predict operation tags per character, the number of characters that need to be generated sequentially, and fill in missing characters. Experiments with it highlight our dataset's complexity, as errors are not simply correctable by basic operations. Moreover, a high PPL score indicates the results lack fluency in LMs' view.

Model	P(%)	R(%)	F_1(%)	Acc(%)	QWK
ChatGPT$_{1-shot}$	50.41	38.38	38.09	44.37	0.1650
ChatGPT$^{\sharp}_{1-shot}$	43.06	41.21	40.34	45.70	0.1933
ChatGLM	47.62	42.32	40.62	46.61	0.2150
ChatGLM$^{\sharp}$	**59.34**	**44.19**	**44.31**	**47.60**	**0.2533**

Table 8: Comparative performance of different setups for Essay Fluency Grading. $^{\sharp}$ indicates the use of all the fine-grained information we annotated.

ChatGPT without finetuning indicated poor rewriting performance, with a large edit distance from the original sentence, as it may generate excessively ornate sentences by rewriting the correct vocabulary or clauses from the original sentence. Such modifications may exceed current students' knowledge and hinder their recognition of issues in their writing. Furthermore, fine-tuning Baichuan on this task achieved the best performance, demonstrating the powerful language understanding and generation capabilities of LLMs. This also underscores the importance of fine-tuning for downstream tasks.

5 Discussion

We explored the importance of fine-grained annotations. Specifically, we studied the significance of grammatical errors for fluency grading and the mutually beneficial relationship between grammatical error types and corrections. Additionally, we also discussed the significance of studying cross-sentence errors.

5.1 Impact of Fine-grained Annotations on Essay Fluency Grading

Leveraging the powerful language understanding capabilities of LLMs, we feed detailed annotations, such as types and counts of errors, into unfinetuned LLMs for the task of `Essay Fluency Grading`. Table 8 shows that fine-grained annotations notably improved performance. Particularly, they improved all metrics for the tunable ChatGLM, and notably increased ChatGPT's recall by 2.83%, confirming the benefits of detailed annotation.

5.2 Max Mutual Benefit of Error Type Identification and Correction

We investigate the mutual benefits between error types and corrections through an explicit prompting approach. Specifically, for `Error Type Identification`, we feed the corresponding corrected sentence along with the input sentence into

Model	Coarse-grained		Fine-grained	
	Micro F$_1$	Macro F$_1$	Micro F$_1$	Macro F$_1$
BERT	69.58	31.29	54.84	15.14
BERT$^\heartsuit$	69.90	28.28	51.07	15.60
BERT$^\spadesuit$	**84.85**	**57.27**	79.71	41.56
RoBERTa	70.34	34.75	56.16	18.63
RoBERTa$^\heartsuit$	70.14	26.80	53.53	15.32
RoBERTa$^\spadesuit$	84.08	54.98	**82.03**	**43.74**

Table 9: A comparison of performance on coarse and fine-grained error type identification with correction reference as inputs. $^\heartsuit$ and $^\spadesuit$ indicate the result after using the silver and gold correction reference.

the model to guide the identification. For `Wrong Sentence Rewriting`, we take the corresponding error types as prompts, feed it into the model, and guide the model to generate the correction for the corresponding error. As ground-truth grammatical error types and corrections are not always available, we also utilize predictions from existing models as inputs (called silver inputs). Specifically, we employ the finetuned RoBERTa from Table 6 to predict grammatical error types and the fintuned BART from Table 7 to generate corrected sentences.

5.2.1 Benefits for Error Type Identification

Table 9 reveals a substantial improvement in error identification, with at least 15% increase in coarse-grained errors and 24% increase in fine-grained errors, when including sentences with ground truth corrections. This emphasizes the effectiveness of utilizing gold corrected sentences for this task and further validates the importance of joint research on grammatical error types and error correction.

Explicitly incorporating predicted corrected sentences (silver inputs) resulted in an average decrease of approximately 2.5% in total compared to the baseline. This decline is attributed to introduced noise, causing the model to learn incorrect relationships between error and corrected sentences. In other words, utilizing more accurate corrected sentences will greatly facilitate the identification of error types, validating the strong correlation between grammatical error types and corrections.

5.2.2 Benefits for Corrections

Table 10 demonstrates a 2% performance increase with models using ground-truth error types. Analysis of coarse versus fine-grained error types revealed that the latter, due to clearer definitions, significantly enhanced correction effectiveness, unlike the negligible impact of coarse-grained types.

Model	EM	F$_{0.5}$	BLEU-4	BERTScore	LD	PPL
BART	18.08	41.21	90.25	97.84	1.67	3.03
BART$^\heartsuit$	17.31	41.49	90.27	97.89	1.43	2.99
BART$^\clubsuit$	18.24	41.80	**90.54**	97.91	**1.48**	**2.97**
BART$^\diamondsuit$	**20.71**	43.00	90.47	**97.94**	1.68	2.98
BART†	19.32	**43.05**	90.18	97.93	1.52	2.98

Table 10: Results on Wrong Sentence Rewriting task with error type as input. $^\heartsuit$ indicate the result after using the silver fine-grained error types. $^\clubsuit$ and $^\diamondsuit$ denotes the model that incorporates the gold coarse and fine-grained error type into the input, while † represents both being used as inputs.

Sent Num	1	2	3	4
Micro F1	32.71	36.30	35.89	**36.88**
Macro F1	11.93	12.22	12.32	**12.53**

Table 11: Results of multi-sentence input on fine-grained error type identification. The columns represent the number of input sentences.

We also conducted a comparison using predicted instead of ground-truth fine-grained error types as input. Compared to BART baseline, there is minimal change across various metrics, but a significant disparity exists compared to models with ground-truth fine-grained error types as input. This indicates that the noise present in the predictions of existing identification models adversely affects the correction model. More accurate and precise grammatical errors will contribute significantly to the correction process.

5.3 Cross-sentence Error

To assess the impact of cross-sentence information on grammar error type identification, we trialed a method increasing input sequence length, shifting from single to multi-sentence recognition. We split and recombine the error sentences in the test set based on their positions in the original essay, constructing input samples with sentence quantities of 1, 2, 3, and 4. We utilized the fine-tuned RoBERTa model mentioned in Table 6 to predict.

Results are shown in Table 11. We observe that for a well-trained model, performance improves with increasing input sequence length. This indicates that cross-sentence information aids in grammatical error type recognition, underscoring the significance of research on cross-sentence errors.

6 Conclusion

We present CEFA, a comprehensive dataset derived from native Chinese student essays. It captures document-level errors, fluency grading, and fine-grained grammatical error details, advancing the field of automated essay fluency assessment. Through experiments using popular existing models, we have demonstrated the challenging nature of our work. Furthermore, we have validated the importance of fine-grained annotation for fluency rating of compositions and the mutually beneficial relationship between error types and corrections.

Limitation

In this section, we address the limitations of our work. Firstly, grammatical errors are just one of the factors affecting essay fluency. As for other factors, our work is limited to reflecting them through fluency grades of essays, leaving significant room for further research in this area. Additionally, the experiments demonstrate that ground-truth corrected sentences and grammatical error types provide significant benefits for error identification and correction. However, such ground-truth information is not readily available in real assessment scenarios. Therefore, our future research will focus on methods that are not solely reliant on ground-truth information. Furthermore, considering the impact of prompt quality on LLMs, the range of prompts we tested for assessing LLMs performance in our tasks was limited.

Acknowledgements

We appreciate the support from National Natural Science Foundation of China with the Main Research Project on Machine Behavior and Human Machine Collaborated Decision Making Methodology (72192820 & 72192824), Fundamental Research Funds for the Central Universities (2024QKT004), Pudong New Area Science & Technology Development Fund (PKX2021-R05), Science and Technology Commission of Shanghai Municipality (22DZ2229004), and Shanghai Trusted Industry Internet Software Collaborative Innovation Center.

References

Emily Ahn, Fabrizio Morbini, and Andrew Gordon. 2016. Improving fluency in narrative text generation with grammatical transformations and probabilistic parsing. In *Proceedings of the 9th International Natural Language Generation Conference*, pages 70–73.

Yigal Attali and Jill Burstein. 2004. Automated essay scoring with e-rater® v. 2.0. *ETS Research Report Series*, 2004(2):i–21.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2022. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, pages 1–59.

Shamil Chollampatt, Weiqi Wang, and Hwee Tou Ng. 2019. Cross-sentence grammatical error correction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 435–445.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Jiefu Gong, Xiao Hu, Wei Song, Ruiji Fu, Zhichao Sheng, Bo Zhu, Shijin Wang, and Ting Liu. 2021. Iflyea: A chinese essay assessment system with automated rating, review generation, and recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 240–248.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! *arXiv preprint arXiv:1809.08731*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Farjana Sultana Mim, Naoya Inoue, Paul Reisert, Hiroki Ouchi, and Kentaro Inui. 2021. Corruption is not all bad: Incorporating discourse structure into pretraining via corruption for essay scoring. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2202–2215.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. Gector–grammatical error correction: tag, not rewrite. *arXiv preprint arXiv:2005.12592*.

OpenAI. 2022. Chatgpt.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, 4:169–182.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Hang Yan, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Chung-Ting Tsai, Jhih-Jie Chen, Ching-Yu Yang, and Jason S Chang. 2020. Lingglewrite: a coaching system for essay writing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 127–133.

Sophie Vanbelle. 2016. A new interpretation of the weighted kappa coefficients. *Psychometrika*, 81(2):399–410.

Baoxin Wang, Xingyi Duan, Dayong Wu, Wanxiang Che, Zhigang Chen, and Guoping Hu. 2022. Cctc: A cross-sentence chinese text correction dataset for native speakers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3331–3341.

Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. A comprehensive survey of grammatical error correction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–51.

Hongyi Wu, Xinshu Shen, Man Lan, Shaoguang Mao, Xiaopeng Bai, and Yuanbin Wu. 2023. A multi-task dataset for assessing discourse coherence in chinese essays: Structure, theme, and logic analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6673–6688.

Lvxiaowei Xu, Jianwang Wu, Jiawei Peng, Jiayu Fu, and Ming Cai. 2022. Fcgec: Fine-grained corpus for chinese grammatical error correction. *arXiv preprint arXiv:2210.12364*.

Min Chul Yang, Min Jeong Kim, Hyoung Gyu Lee, and Hae Chang Rim. 2012. Assessing writing fluency of non-english-speaking student for automated essay scoring: How to automatically evaluate the fluency in english essay. In *4th International Conference on Computer Supported Education, CSEDU 2012*, pages 83–87.

Yiqin Yang, Li Xia, and Qianchuan Zhao. 2019. An automated grader for chinese essay combining shallow and deep semantic attributes. *IEEE Access*, 7:176306–176316.

Zheng Yuan and Christopher Bryant. 2021. Document-level grammatical error correction. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 75–84.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022a. Mucgec: a multi-reference multi-source evaluation dataset for chinese grammatical error correction. *arXiv preprint arXiv:2204.10994*.

Yue Zhang, Bo Zhang, Haochen Jiang, Zhenghua Li, Chen Li, Fei Huang, and Min Zhang. 2023. Nasgec: a multi-domain chinese grammatical error correction dataset from native speaker texts. *arXiv preprint arXiv:2305.16023*.

Zhexin Zhang, Jian Guan, Guowei Xu, Yixiang Tian, and Minlie Huang. 2022b. Automatic comment generation for chinese student narrative essays. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 214–223.

Honghong Zhao, Baoxin Wang, Dayong Wu, Wanxiang Che, Zhigang Chen, and Shijin Wang. 2022. Overview of ctc 2021: Chinese text correction for native speakers. *arXiv preprint arXiv:2208.05681*.

Xinlin Zhuang, Hongyi Wu, Xinshu Shen, Peimin Yu, Gaowei Yi, Xinhao Chen, Tu Hu, Yang Chen, Yupei Ren, Yadong Zhang, Youqi Song, Binxuan Liu, and Man Lan. 2024. TOREE: Evaluating topic relevance of student essays for Chinese primary and middle school education. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5749–5765, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

A Annotation Specification

A.1 Essay Fluency Grading

Essay fluency grading adheres to the following criteria:

- Excellent (2 points): The types of grammatical errors committed do not affect reading fluency (e.g., Typographical Error and Factual Illogicality). The annotator, when reading through once, encounters no stumbling or incomprehensible parts.
- Average (1 point): The types of grammatical errors affecting reading fluency (the other 16 types of errors) do not exceed five sentences. The annotator, when reading through once, stumbles or finds parts hard to understand no more than five times.
- Unsatisfactory (0 points): The types of grammatical errors affecting reading fluency (the other 16 types of errors) exceed five sentences. The annotator, when reading through once, stumbles or finds parts hard to understand more than five times.

A.2 Error Types

After conducting in-depth research into primary and secondary school student writing and extensively investigating the development of GEC data annotation standards, we have re-examined the classification of grammar errors in GEC and synthesized a revised set of annotation standards. Our annotation specification holistically covers simple grammatical errors such as punctuation and spelling mistakes, as well as complex grammatical issues like missing components and improper collocations, offering a further categorization of grammar errors and corresponding correction methods. Specifically, in terms of grammar error types, we have classified the grammatical errors in compositions into character-level and component-level errors, further divided into 5 coarse-grained and 18 fine-grained error types. Our annotations adhere to the principle of minimal modification. Our newly summarized definitions for grammatical error types are as follows:

Character-Level Error (CL). Including four fine-grained error types: Word Missing (**WM**), where a word in a commonly used fixed collocation is missing from the sentence and needs to be added; Typographical Error (**TE**), where there are typos in the sentence that need to be revised or deleted; Missing Punctuation (**MP**), where punctuation is missing from the sentence and needs to be added; and Wrong Punctuation (**WP**), where the punctuation used in the sentence is wrong and needs to be revised or deleted.

Redundant Component Error (RC). Four fine-grained error types are: Subject Redundancy (**SR**), which occurs when a complex adverb immediately follows the first subject, followed by another subject referring to the same thing, and the modification is to delete one subject; Particle Redundancy (**PR**) refers to the redundant use of particles, which should be deleted during editing. Statement RePetition (**SRP**) occurs when some words or clauses repeat in the sentence, and the solution is to delete them. Other Redundancy (**OR**) refers to any redundant elements not covered by the previous types, which should also be deleted in modification.

Incomplete Component Error (IC). Four fine-grained error types with missing components are: Unknown Subject (**US**), which occurs when the sentence lacks a subject or the subject is unclear, and the solution is to add or clarify the subject; Predicate Missing (**PM**) refers to a sentence lacking verbs, which can be corrected by adding predicates; OBject Missing (**OBM**) means that a sentence lacks an object, and the solution is to add an object; OTher Missing (**OTM**) refers to other missing components besides the incomplete subject, predicate, and object, which can be corrected by adding the missing components except for the subject, predicate, and object.

Incorrect Constituent Combination Error (ICC). Including four fine-grained error types: Inappropriate Subject-Verb Collocation (**ISVC**), which occurs when the subject and predicate are not properly matched, and can be corrected by replacing either the subject or predicate with other words. Inappropriate Verb-Object Collocation (**IVOC**) refers to the predicate and object not being properly matched, and can be corrected by replacing either the predicate or object with other words. Inappropriate Word Order (**IWO**) means that the order of words or clauses in the sentence is unreasonable, and can be corrected by rearranging some words or clauses. Inappropriate Other Collocation (**IOC**) refers to any element in the sentence not covered by the previous types being improperly matched, and can be corrected by replacing it with other words.

Illogical (IL). This includes two subcategories:

Type	Example
SRD	**Sent:** 我在阳台上一共种了两株，我平时见不到它们。 (I planted a total of two on the balcony, I usually don't see them.) **Tips:** Delete the second subject, "I".
PR	**Sent:** 由于邓稼先的癌症的越来越严重，经常病地倒在了地上。 (As Deng Jiaxian's cancer became more serious, he often fell ill to the ground.) **Tips:**Delete the second "的".
SRP	**Sent:** 数字又不只是一个数字，在这个快速发展的时代里，我们每天都可以看到不同的数字，可其中的它们又不是一个数字，因为背后都是真实发生的事。 (Number is not just number. In this era of rapid development, we can see different numbers every day, but they are not just numbers, as behind them are real events.) **Tips:** "Number is not just number" repeats with "they are not just numbers".
OR	**Sent:** 一个易拉罐被踢开了下山去。 (A soda can was kicked away and went down the hill.) **Tips:** "kicked away and went down the hill" equals to "kicked down the hill"
US	**Sent:** 眼泪瞬间流下，滴落在了衣服上，出现深色小圆点，又接二连三的掉下来。 (Tears flowed down in an instant, dripping onto the clothes, small dark dots appeared, and fell down one after another.) **Tips:** Subjects changed in clauses. Add subject "tears" before "fell down".
PM	**Sent:** 邓稼先从美国后，就立刻接到了研究原子弹工作。 (After Deng came from US, he at once received a job to study the atomic bomb.) **Tips:** Add "归来" after "美国".
OBM	**Sent:** 然而我想说，并不是所有书籍都有能力完成承载读者。 (However, I want to say that not all books are capable of carrying readers.) **Tips:** Add "任务" after "承载读者".
OTM	**Sent:** 爱迪生为改良电灯试用6000多材料，试验7000多次。 (Edison tried over 6000 materials and over 7000 tests to improve the electric lamp.) **Tips:** Add "种" after "6000多".
ISVC	**Sent:** 他知道我们比较薄弱的地方，并使我们在下一次测试中得到提高。 (He knows where we are weak and improves us for the next test.) **Tips:** Predicate "提高" should be paired with subject "我们的成绩", not "我们".
IVOC	**Sent:** 我尽管不是班里最高分，但也达到了很大的进步。 (Although I am not the highest score in the class, I have made great progress.) **Tips:** Object "进步" should be paired with predicate "取得" instead of "达到".
IWO	**Sent:** 一次受到生活打击的祥子也没有放弃。 (Xiangzi who was hit by life once did not give up.) **Tips:** "一次" should be placed after "祥子".
IOC	**Sent:** 牛顿被苹果为什么会从树下掉下来感到困惑，最后研究出了万有引力定律。 (Newton was puzzled by why the apple fell from the tree, and finally worked out the law of gravitation.) **Tips:** "感到困惑" should be paired with "为" instead of "被".
FIL	**Sent:** 聂海胜出生在湖北枣庄一个物质极度匮乏的小山村中。 (Nie Haisheng was born in a small mountain village in Zaozhuang, Hubei, where materials are extremely scarce.) **Tips:** Nie Haisheng was born in Zaoyang, Hubei, not in Zaozhuang, Hubei.
LIL	**Sent:** 那老奶奶抬起头，只是一惊，然后便笑着说：“没事，谢谢小伙子的好心，我自己来就好。” (The old woman raised her head, was just surprised, and then said with a smile: "It's okay, thank you for your kindness, I'll just do it myself.) **Tips:** The action 'surprised' comes before 'smiling.' When describing 'being surprised,' we should use "先是"(firstly) rather than "只是"(just).

Table 12: Examples of each fine-grained component-level error types.

Factual Illogicality (**FIL**) and Linguistic Illogicality (**LIL**). The former refers to instances that conflict with factual information, while the latter refers to misuse of logical conjunctions, idioms, etc., that render the sentence illogically constructed.

Table 12 shows examples of each fine-grained error type.

B Inter-Annotator Agreement (IAA) Calculation

In this study, we adopted an Inter-Annotator Agreement (IAA) measure. For the `Error Type Identification` and `Essay Fluency Grading` tasks, we employed Cohen's Kappa to measure the consistency among annotators. For the `Wrong Sentence Rewriting` task, we used the $F_{0.5}$ score for the same purpose. The annotation was divided into five batches, with each batch containing 100, 100, 60, 80, and 161 essays, respectively. The consistency scores for each batch detailed in the corresponding Table 13.

Task	Batch 0	Batch 1	Batch 2	Batch 3	Batch 4	Avg.
Error Types	0.6906	0.5504	0.5493	0.5291	0.6133	0.5865
Correction	78.65%	57.71%	59.05%	51.56%	63.64%	62.12%
Grading	0.6628	0.5846	0.5938	0.5586	0.6184	0.6036

Table 13: The consistency analysis results demonstrate the IAA scores, represented as percentages, across various aspects of text analysis for different data sub-batches (each batch representing a round of annotation). The final column indicates the average annotator consistency score across all batches.

C Detailed results for Error Type Identification

Table 14 presents the model's identification results for various coarse-grained error types. Clearly, the model demonstrates better learning and prediction performance for error types with higher frequencies (CL). However, for error types with lower frequencies (ICC, IL), the model struggles to learn their distinctive features, resulting in poorer prediction performance.

D Prompt for Models

We have listed the prompts used for all tasks, including `Essay Fluency Grading`, `Error Type Identification` and `Wrong Sentence Rewriting`. Note that the original prompts were written in Chinese, and we provide their English translations here.

D.1 Essay Fluency Grading

The prompts we use for this task are as follows:

Zero-shot prompt for ChatGPT, where [E] is the essay:

> "Assuming you are a primary or secondary school language instructor, I will provide you with an essay. Please evaluate its fluency on a scale of 0 to 2: where 0 denotes "Not Fluent", 1 denotes "Moderately Fluent", and 2 denotes "Highly Fluent". Kindly return only the fluency score. Input: [E]; Output:"

Few-shot prompt for ChatGPT, where [E] is the essay, and [G] is the fluency grade of [E].:

> "Assuming you are a primary or secondary school language instructor, I will

Model	CL	RC	IC	ICC	IL	Micro F_1	Macro F_1	Micro			Macro		
								P	R	F_1	P	R	F_1
FCGEC	88.97	**25.43**	31.33	2.82	0.00	69.25	29.71	38.88	**53.12**	44.90	9.48	13.33	9.52
BERT	87.93	20.00	40.74	7.79	0.00	69.58	31.29	67.18	46.33	54.84	18.68	13.54	15.14
RoBERTa	88.51	25.00	**46.23**	14.00	0.00	**70.34**	**34.75**	66.67	48.51	**56.16**	**22.84**	16.51	**18.63**
ChatGPT$_{0-shot}$	16.93	21.50	12.79	14.06	0.00	15.41	13.05	8.58	13.26	10.42	9.45	**17.31**	7.27
ChatGPT$_{3-shot}$	44.64	21.82	4.35	12.21	1.80	25.49	16.96	11.25	13.82	12.40	12.25	14.50	8.51
ChatGLM$_{0-shot}$	0.38	12.99	21.37	0.00	0.47	5.30	7.04	5.09	4.68	4.87	7.18	9.53	4.92
ChatGLM$_{3-shot}$	16.10	25.93	12.57	0.00	0.45	14.91	11.01	5.58	4.99	5.27	11.81	7.67	3.57
ChatGLM$_{ft}$	**89.26**	24.73	26.25	**16.49**	0.00	67.75	31.35	52.04	47.06	**49.42**	18.60	14.63	15.50
Baichuan$_{ft}$	87.22	24.66	35.81	2.38	0.00	65.61	30.01	52.42	49.42	50.88	13.52	12.96	13.13
BERT$^{\heartsuit}$	88.25	13.53	31.11	8.51	0.00	69.90	28.28	62.56	43.15	51.07	22.11	13.19	15.60
RoBERTa$^{\heartsuit}$	88.56	12.70	27.91	4.82	0.00	70.14	26.80	**67.59**	44.31	53.53	21.67	12.89	15.32

Table 14: Comparison of performance on coarse and fine-grained error type identification. The left is the results of coarse-grained error type identification. The right is the fine-grained one.

> provide you with an essay. Please evaluate its fluency on a scale of 0 to 2: where 0 denotes "Not Fluent", 1 denotes "Moderately Fluent", and 2 denotes "Highly Fluent". Kindly return only the fluency score. Here are some samples: Sample 1: Input: [E]; Output: [G]. Input: [E]; Output:"

Prompts for ChatGLM and Baichuan is the same as zero-shot prompt for ChatGPT.

D.2 Error Type Identification

Zero-shot prompt for ChatGPT in both coarse-grained and fine-grained error type identification, where [S] indicates the sentence:

> "Assume you are a primary or secondary school language instructor proficient in grammar type identification and correction for student essays. In this context, I have defined five error categories. I will list these categories in the format "Error Type ID, Error Type: Definition;". Please identify the error types in the given sentence. Note that a sentence might contain multiple error categories. Kindly return the identification and correction results in the JSON format: "errorTypeId":[Error Type ID$_1$, Error Type ID$_2$], "errorType":[Error Type 1, Error Type 2], "revisedSent":"Corrected Sentence". If you believe the sentence is grammatically correct, please return "errorTypeId":[0], "errorType":["Right"]. The definitions are as follows: [Error Type ID], [Error Type]: [Definition]; Input: [S]; Output:"

Few-shot prompt for ChatGPT in both coarse-grained and fine-grained error type identification, where [S] indicates the sentence and [E] denotes the error type:

> "Assume you are a primary or secondary school language instructor proficient in grammar type identification and correction for student essays. In this context, I have defined five error categories. I will list these categories in the format "Error Type ID, Error Type: Definition;". Please identify the error types in the given sentence. Note that a sentence might contain multiple error categories. Kindly return the identification and correction results in the JSON format: "errorTypeId":[Error Type ID$_1$, Error Type ID$_2$], "errorType":[Error Type 1, Error Type 2], "revisedSent":"Corrected Sentence". If you believe the sentence is grammatically correct, please return "errorTypeId":[0], "errorType":["Right"]. The definitions are as follows: [Error Type ID], [Error Type]: [Definition]. Here are some samples: Input: [S], Output: "errorTypeId":[1,2], "errorType":[[E$_1$], [E$_2$]] Input: [S]; Output:"

Similarly, prompts for ChatGLM and Baichuan is the same as zero-shot prompt for ChatGPT.

Specifically, our input prompt augmented with revised sentence is as follows, where [S] denotes the original sentence and [R] represents the revised sentence:

> "Assume you are a primary or secondary school language instructor proficient in grammar type identification for student essays. In this context, I have defined five error categories. I will list these categories in the format "Error Type ID, Error Type: Definition;". Please identify the error types in the given sentence and revised sentence. Note that a sentence might contain multiple error categories. Kindly return the identification and correction results in the JSON format: "errorTypeId":[Error Type ID$_1$, Error Type ID$_2$], "errorType":[Error Type 1, Error Type 2], "revisedSent":"Corrected Sentence". If you believe the sentence is grammatically correct, please return "errorTypeId":[0], "errorType":["Right"]. The definitions are as follows: [Error Type ID], [Error Type]: [Definition]. Sentence: [S]; Revised Sentence: [R]; Output: "

D.3 Wrong Sentence Rewriting

Zero-shot prompt for ChatGPT, where [S] denotes the wrong sentence:

> “You are an elementary or secondary school language teacher tasked with correcting erroneous sentences in student essays. I will provide you with a sentence from the essay; please make necessary revisions. Bear in mind, adjustments should adhere to the principle of minimal change. Kindly return only the revised sentence. If you believe the sentence is error-free, simply return the input sentence. Input: [S]; Output:”

Few-shot prompt for ChatGPT, where [S] denotes the wrong sentence and [R] indicates the revised sentence:

> “You are an elementary or secondary school language teacher tasked with correcting erroneous sentences in student essays. I will provide you with a sentence from the essay; please make necessary revisions. Bear in mind, adjustments should adhere to the principle of minimal change. Kindly return only the revised sentence. If you believe the sentence is error-free, simply return the input sentence. Input: [S]; Output: [R]; Input: [S]; Output:”

Similarly, prompts for ChatGLM and Baichuan is the same as zero-shot prompt for ChatGPT.

Specifically, our input prompt augmented with error type information is as follows, where [S] indicates the sentence and [E] denotes the error types:

> "You are a primary and secondary school language teacher capable of correcting erroneous sentences from student essays. I will provide you with a sentence from the essay along with its error category. Please make corrections based on the provided error category, adhering to the principle of minimal changes. Only return the revised sentence; if you believe the sentence is error-free, return the original sentence. I will list these categories in the format "Error Type ID, Error Type: Definition;". The definitions are as follows: "[Error Type ID], [Error Type]: [Definition];" Sentence: [S]; Error Type: [E]; Output: "