FEANEL: A Benchmark for Fine-Grained Error Analysis in K-12 English Writing

Jingheng Ye[1,2], **Shen Wang**[1], **Jiaqi Chen**[2], **Hebin Wang**[2],
Deqing Zou[2], **Yanyu Zhu**[2], **Jiwei Tang**[2], **Hai-Tao Zheng**[2],
Ruitong Liu[2], **Haoyang Li**[1], **Yanfeng Wang**[3], **Qingsong Wen**[1]
[1]Squirrel Ai Learning, [2]Tsinghua University, [3]Shanghai Jiao Tong University
yejh22@mails.tsinghua.edu.cn

Abstract

Large Language Models (LLMs) have transformed artificial intelligence, offering profound opportunities for educational applications. However, their ability to provide fine-grained educational feedback for K-12 English writing remains underexplored. In this paper, we challenge the error analysis and pedagogical skills of LLMs by introducing the problem of Fine-grained Error Analysis for English Learners and present the *Fine-grained Error ANalysis for English Learners* (FEANEL) Benchmark. The benchmark comprises 1,000 essays written by elementary and secondary school students, and a well-developed English writing error taxonomy. Each error is annotated by language education experts and categorized by type, severity, and explanatory feedback, using a part-of-speech-based taxonomy they co-developed. We evaluate state-of-the-art LLMs on the FEANEL Benchmark to explore their error analysis and pedagogical abilities. Experimental results reveal significant gaps in current LLMs' ability to perform fine-grained error analysis, highlighting the need for advancements in particular methods for educational applications.

1 Introduction

Large Language Models (LLMs) have revolutionized artificial intelligence with their extensive knowledge and remarkable reasoning capabilities [21], creating unprecedented opportunities in educational applications [55, 57, 8, 56]. In language education, LLM-powered solutions are increasingly being deployed to enhance personalized learning experiences [63, 37]. However, while LLMs demonstrate impressive performance in many tasks, their application in providing fine-grained educational feedback targeted at each error students may make [54, 48, 58, 19], which is critical for language acquisition, remains under-explored.

Current related methodologies, however, typically focus on surface-level corrections [5, 22, 60] or global assessments with coarse feedback [12, 27], which do not capture the multifaceted nature of writing difficulties. Moreover, the lack of a standardized taxonomy for English writing errors [68] has led to inconsistencies in error categorization and hindered the development of robust educational tools. These gaps are particularly pronounced in K-12 English education, where learners exhibit diverse proficiency levels and error patterns that require fine-grained analysis and personalized feedback.

Therefore, we define the problem of *Fine-grained Error Analysis for Language Learners*, a crucial component of language education aimed at systematically analyzing learners' errors in written English. Error analysis, as a foundational methodology in second language acquisition research [26, 14], serves two primary purposes: (1) investigating the underlying causes of errors to facilitate targeted interventions, and (2) providing insights into common difficulties in language learning to inform teaching practices and materials. By offering detailed feedback on error types, severity, and possible

Preprint. Under review.

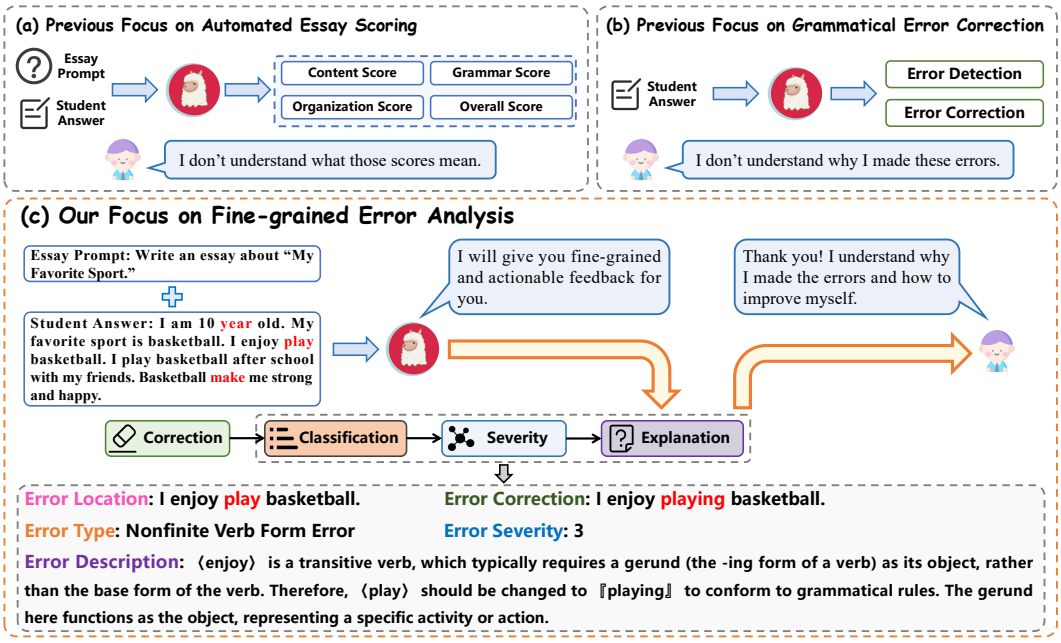


Figure 1: Comparison of our focus on Fine-grained Error Analysis with existing studies.

corrections, this problem not only supports learners in scaffolding their knowledge but also enhances their ability to learn from mistakes through instant and interpretable feedback [10, 62].

To investigate this problem, we introduce the *Fine-grained Error ANalysis for English Learners* (FEANEL) Benchmark, designed to advance research in fine-grained error analysis. The benchmark includes a large-scale dataset of 1,000 essays written by K-12 students, with 500 essays from elementary school students and 500 from secondary school students, covering a wide range of age groups and proficiency levels. Each error analysis has been meticulously annotated with an error type, severity level, and explanation, guided by a taxonomy co-developed with language education experts.

As illustrated in Figure 1, this paper goes beyond conventional automated essay scoring [27] and grammatical error correction [5, 65] by highlighting the interpretability and educational value of feedback, thereby facilitating a more thorough evaluation of LLMs within language education. Moreover, the benchmark establishes a rigorous framework for evaluating fundamental competencies of LLMs, including their comprehension of syntactic, grammatical, and lexical knowledge and their capacity to replicate pedagogical scenarios by producing engaging and didactically significant feedback. Additionally, FEANEL also evaluates LLMs' capacity for commonsense reasoning and knowledge application, recognizing that effective error analysis in essays often involves understanding logical relationships and world knowledge (e.g., an essay requiring an introduction of tourist attractions in Beijing). This multidimensional evaluation provides a more detailed understanding of LLMs' efficacy in providing insightful and contextually relevant feedback to learners.

We conduct extensive experiments to evaluate the performance of various LLMs on FEANEL. Our empirical study reveals: (1) LLMs still face significant challenges in classifying complex errors, and often fall short of human-level pedagogical nuance in their explanations. (2) Performance of LLMs is highly dependent on the detail level of prompts and the availability of examples. (3) LLMs are sensitive to the sub-task execution order. We believe that our proposed FEANEL and findings are crucial for educational applications and understanding LLMs' pedagogical ability. In summary, our contributions are threefold:

- We define the problem of Fine-grained Error Analysis and introduce the FEANEL Benchmark, a novel dataset annotated by English education experts for fine-grained error analysis in writing.
- We develop a well-defined and part-of-speech-driven taxonomy for English writing errors, addressing the issues of inconsistent categorization and insufficient granularity in previous work.

- We conduct a comprehensive empirical evaluation of various LLMs, providing insights into their capabilities and limitations in generating interpretable and pedagogically valuable feedback.

2 The FEANEL Benchmark

2.1 Problem Definition

Given an essay prompt P, a student's written answer X, its corrected version Y, and a set of edits $\mathbb{E} = \{e_1, e_2, \cdots, e_N\}$ that transform X to Y[1], the problem of fine-grained error analysis focuses on analyzing each specific edit/error. Each analysis comprises three key elements: (1) **Error Classification**: Categorize the error into an error type $t_i \in \mathcal{T}$ based on the pre-defined taxonomy $\mathcal{T}$. (2) **Error Severity Rating**: Assign a numerical score $s_i \in \{1, 2, 3, 4, 5\}$ to indicate how critically each error affects the sentence's overall structure and meaning. (3) **Error Explanation**: Provide an accurate, relevant, and sufficient explanation d_i of why the error occurred and how to correct or prevent it. Notice that there may be multiple errors in an edit. We require LLMs to assign a single error type with the highest priority, yet explain all error types in the explanation. By default, LLMs are required to generate an error analysis in the order of Error Severity $\rightarrow$ Error Type $\rightarrow$ Error Explanation, which is formulated as:

$$P(s_i, t_i, d_i \mid X, Y, e_i) = P(s_i \mid X, Y, e_i) \cdot P(t_i \mid X, Y, e_i, s_i) \cdot P(d_i \mid X, Y, e_i, s_i, t_i) \quad (1)$$

2.2 Dataset Construction.

Data Collection. We collected 1,000 original essays from two distinct sources to ensure diversity in learner proficiency and educational context. First, 500 essays from elementary school students were collected via a global online education platform. The prompts were specifically designed to elicit particular *tenses*, thereby providing clear instances for analyzing grammatical understanding. Another 500 essays of school students were sourced from the TECCL Corpus[2], a significant corpus of Chinese EFL learners' writing. It is notable for its wide array of over 1,000 essay topics and its representation of learners from elementary to postgraduate levels. For our study, we selected essays corresponding to middle school proficiency. All data from both sources were fully anonymized before use to ensure privacy compliance. This combined dataset offers a rich spectrum of writing capabilities for fine-grained error analysis.

Data Cleaning. To ensure the quality and relevance of the data for our analysis, a rigorous cleaning process was implemented by a team of annotators. Essays were excluded if they: (1) significantly deviated from the given prompt, rendering them off-topic; (2) contained an insufficient number of words, indicating a lack of substantive response; (3) were entirely error-free, as our focus is on error analysis; or (4) were incomplete or nonsensical. Following this filtering, the original formatting of the retained essays was preserved to maintain the authenticity of student responses with real errors.

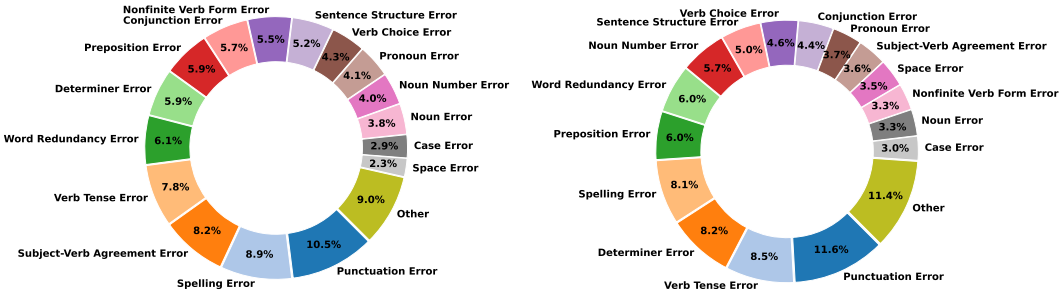
Data Annotation. The data underwent annotation by English education experts. The comprehensive annotation workflow was as follows:

(1) **Error Detection and Correction**: Following established Grammatical Error Correction (GEC) guidelines [5], education experts rewrote student essays, applying the *minimal* necessary corrections to preserve the original meaning. This principle ensures objectivity and facilitates accurate error categorization. Experts subsequently reviewed these corrections, checking for over-corrections, missed errors, or incorrect revisions to guarantee accuracy and necessity. We then leverage the edit extraction tool *CLEME* [60, 61] to extract a set of edits describing the revision trajectory from X to Y for subsequent annotation of error analysis.

(2) **Error Type Taxonomy**: In collaboration with two experienced educators, we developed an error type taxonomy comprising 29 distinct error types, primarily based on part-of-speech categories (see Appendix B). The taxonomy was designed for *broad coverage* and *mutual exclusivity* with appropriate granularity, enabling precise classification of nearly all errors without resorting to a

[1] In this paper, each edit $e_i \in \mathbb{E}$ is considered an instance of an error, as edits are introduced only to rectify mistakes.

[2] https://corpus.bfsu.edu.cn/info/1070/1449.htm



(a) Error type distribution of elementary essays.

(b) Error type distribution of secondary essays.

Figure 2: Error type distribution of FEANEL. We illustrate the most frequent 16 error types due to the space limit and present the remaining error types as "Other".

vague "Other Error" category. A prioritization system was established in Appendix B to identify the salient error type for complex errors involving multiple error types.

(3) **Error Analysis**: Each identified error underwent a three-step analysis: (1) *Error Classification*: Experts assigned the most prioritized error type to each edit. (2) *Error Severity Rating*: A 1-5 scale was used to determine the seriousness of each error's impact on sentence structure and meaning. (3) *Error Explanation*: Two experts independently provided detailed explanations for each error, adhering to principles of accuracy, relevance, and sufficiency. Another expert then reviewed and selected the most appropriate explanation, refining it for clarity and completeness if necessary.

2.3 Dataset Analysis

Dataset Statistics. The overall statistics of the FEANEL benchmark are presented in Table 1. The dataset comprises 1,000 essays, evenly split with 500 from elementary school students and 500 from secondary school students. Secondary school essays are considerably longer. This difference in length correlates with the number of identified edits: elementary school essays contain 3,005 edits, while secondary school essays feature significantly more, 5,671 edits. In total, the benchmark includes 8,676 fine-grained error analyses. Interestingly, the average length of an edit is slightly higher for elementary students (1.66 words) than for secondary students (1.53 words). This may suggest that errors made by younger learners with lower language proficiency can be more substantial or involve more words.

Table 1: Dataset statistics of the FEANEL benchmark.

	Essays	Essay Len.	Edits/Essay	Edit Len.	Exp. Len.
Ele.	500	48.2	6.01	1.66	69.27
Sec.	500	127.1	11.34	1.53	52.79

Error type Distribution. The distribution of error types across elementary and secondary school student essays in Figure 2 reveals the error distributions are quite similar. Key error categories such as Punctuation Error, Spelling Error, Verb Tense Error, Word Redundancy Error, Determiner Error, and Preposition Error each account for over 5% of the total errors observed in both elementary and secondary school essays. These represent persistent challenges for K-12 English language learners. However, notable divergences also exist. For instance, Subject-Verb Agreement errors are proportionally more prevalent in the writing of elementary school students, which is consistent with typical language acquisition trajectories. Essays from secondary school students tend to exhibit a higher proportion of errors classified under the "Other" category. This suggests a more pronounced long-tail effect in their error patterns, possibly due to their engagement with more complex linguistic structures and vocabulary, leading to a wider variety of less common errors.

2.4 Evaluation Metrics.

We evaluate the fine-grained error analysis task along its three core components. For Error Classification, we report Accuracy to capture overall correctness and Macro-F1 to ensure balanced assessment across all categories, particularly highlighting performance on long-tail distributions. Error Severity Rating is assessed using Mean Absolute Error (MAE). To evaluate Error Explanation quality, we

employ standard n-gram-based metrics, i.e., BLEU [40], METEOR [2], and ROUGE-L [34]. This combined set of metrics provides a robust and complementary evaluation protocol.

3 Experiments

3.1 Experimental Settings

Baseline Models. We evaluate the performance of various LLMs on our FEANEL benchmark to investigate their ability to perform error analysis. The experiment involve state-of-the-art reasoning models as well as other representative models, including GPT-4o [24], o1 [25], o3 [39], o4-mini [39], Gemini-2.5-pro [50], DeepSeek-R1 [18], Claude-3.7-Sonnet [9], Claude-3.7-Sonnet-Thinking [9], Grok-3-Beta [17], Qwen-3 [51], Llama-3 [16], and Mistral-Small-3.1 [1]. These models represent a mix of closed-source commercial systems and open-source models, ensuring a comprehensive evaluation across different architectures and training paradigms. We report evaluation details and design prompts in Appendix C.

Evaluation Settings. We design three distinct experimental settings to evaluate the LLMs' performance on the FEANEL benchmark under varying levels of instructional detail. This comprehensive approach allows us to systematically assess the LLMs' intrinsic understanding of the task versus their ability to leverage explicit guidance, providing insights into their robustness and adaptability. (1) `Zero-shot-naive`: LLMs are provided with only the basic task instruction and the label space for our error taxonomy. Crucially, they do not receive any demonstrations, detailed definitions, or illustrative examples for each error type, nor are they given definitions for the severity scores (1-5). The purpose of this setting is to test the LLMs' unassisted ability to perform fine-grained error analysis with minimal contextual information, thereby establishing a baseline for their inherent capabilities. (2) `One-shot-detailed`: The setting builds directly upon the `Zero-shot-naive` setting by offering more comprehensive guidance and a demonstration. Models receive detailed definitions and examples for every error type within our taxonomy, along with explicit definitions for each severity score from 1 to 5. The given demonstration allows us to investigate the LLMs' capacity for in-context learning and their ability to generalize effectively from a specific, relevant example.

The progression from `Zero-shot-naive` to `One-shot-detailed` systematically increases the richness and explicitness of the input prompt. This design gradually reduces the task's inherent ambiguity and difficulty for the models.

3.2 Evaluation Results

The comprehensive results of our experiments across the three evaluation settings are presented in Table 2 for `Zero-shot-naive` and Table 3 for `One-shot-detailed`. Our analysis across these settings reveals several key insights into the capabilities and limitations of current LLMs.

Overview results. The results reveal that no LLM consistently outperforms others across all three sub-tasks. For Error Classification, larger models often designated as thinking models, including Gemini-2.5-pro, o3-low, o1, o4-mini, Claude-3.7-Thinking, and DeepSeek-R1, generally demonstrate superior accuracy. This may be attributed to their enhanced reasoning capabilities, which are beneficial for systematically applying our taxonomy to complex linguistic errors. In the Error Severity Rating task under the `Zero-shot-naive` setting, models such as Grok-3, o1, o3-low, and Claude-3.7-Thinking showed stronger performance, potentially reflecting better intuitive calibration for impact assessment. However, with the provision of detailed definitions and an example in the `One-shot-detailed` setting, models like Claude-3.7, Llama-3.3-70b, and Claude-3.7-Thinking excelled. For the Error Explanation task, Gemini-2.5-pro exhibited a significant lead, followed by models like Claude-3.7, o1, o3-low, and o4-mini. This highlights that models with strong generative capabilities are better suited for producing high-quality and pedagogically relevant textual feedback.

For the error classification task. We observe that classifying errors in essays from elementary school students is generally more challenging than for those from secondary school students. Across various models, the accuracy (Acc) for elementary-level essays is typically 2$\sim$6 percentage points lower than for secondary-level essays. We hypothesize this is primarily because elementary students are more prone to making compound errors, where a single edit may involve multiple intertwined

Table 2: Main results of the `Zero-shot-naive` setting. We color the the secondary results.

Model	Think	Classification				Severity		Explanation					
		Acc↑		F_1↑		MAE↓		BLEU↑		METEOR↑		ROUGE↑	
GPT-4o	✗	61.74	66.79	46.55	52.67	0.87	0.77	1.35	1.19	17.87	17.58	24.29	23.02
o1	✓	68.60	74.27	**62.64**	64.50	0.68	0.60	0.82	1.00	15.37	16.12	23.95	23.04
o3-low	✓	70.93	74.79	62.47	**66.06**	0.69	0.63	1.32	1.16	17.62	17.10	24.12	22.95
o4-mini-low	✓	69.80	71.43	55.28	60.97	0.82	0.80	1.44	1.18	17.73	16.44	26.06	23.40
o4-mini-medium	✓	68.83	70.87	54.63	58.86	0.79	0.80	1.48	1.29	18.07	16.85	26.43	23.78
o4-mini-high	✓	69.56	72.89	55.72	61.33	0.79	0.46	1.51	1.34	18.03	17.50	26.14	24.51
Gemini-2.5	✓	**72.06**	**76.34**	60.67	65.95	0.80	0.79	**3.15**	**2.61**	**25.36**	**24.21**	**28.61**	26.15
DeepSeek-R1	✓	67.87	72.51	54.79	60.41	0.76	0.72	1.57	1.37	17.25	16.83	28.31	**26.29**
Claude-3.7	✗	64.57	70.04	51.56	58.60	0.76	0.68	2.50	2.20	22.17	22.81	26.42	24.90
Claude-3.7	✓	71.40	74.65	58.59	60.56	0.69	0.65	2.17	2.00	21.12	21.41	26.82	25.13
Grok-3	✗	66.90	72.69	52.92	59.14	**0.67**	**0.59**	2.37	1.70	24.67	24.07	26.06	23.55
Mistral-small	✗	57.48	66.03	44.34	50.92	0.75	0.65	1.84	1.56	20.37	20.62	25.58	23.87
Qwen-3-8b	✓	57.44	61.28	41.04	45.38	0.90	0.79	0.97	0.67	15.62	14.74	23.51	21.45
Qwen-3-30b-a3b	✓	61.57	65.19	44.42	49.83	0.76	0.70	0.92	0.81	15.39	15.59	24.49	22.16
Qwen-3-230b-a22b	✓	62.50	68.66	49.84	55.64	0.84	0.82	0.82	0.82	14.42	14.31	22.64	21.23
Llama-3.1-8b	✗	35.33	40.25	23.17	26.07	1.13	1.10	0.64	0.54	15.97	15.54	21.47	19.61
Llama-3.1-70b	✗	53.28	59.12	42.46	45.04	0.79	0.73	1.24	0.83	17.37	16.38	23.42	21.25
Llama-3.3-70b	✗	56.48	63.12	41.84	46.69	0.95	0.92	1.52	1.17	19.28	19.46	21.48	20.26
Average	-	63.13	67.83	50.16	54.92	0.80	0.73	1.54	1.30	18.54	18.2	24.99	23.14
Human	-	79.90	76.66	60.53	62.25	0.99	0.72	5.21	5.20	25.28	28.39	33.50	31.60

linguistic issues, often spanning more words (Table 1). This inherent complexity in the error instances naturally increases the difficulty of assigning a single salient error category. Furthermore, the Macro F1 scores for error classification are consistently and significantly lower than accuracy scores across all models and settings. This discrepancy indicates that while models may perform reasonably well on frequent error types, their ability to accurately classify less frequent error types remains limited. A detailed analysis of model performance on each specific error type is presented in Appendix D.

For the connection between different sub-tasks. A notable trend emerging from our results is that models exhibiting superior performance in the Error Classification task also tend to generate higher-quality explanations. This strong positive correlation suggests an intrinsic link between the ability to correctly categorize an error and the ability to articulate a meaningful and pedagogically sound explanation for it. This finding underscores the importance of understanding accurate errors as a foundational prerequisite for effective feedback generation. To further investigate this phenomenon and the potential benefits of optimizing this interplay, we conduct an ablation study, presented in Section 3.5, where we explore the impact of varying the execution order.

For the impact of the information richness of the prompt. Our experiments demonstrate a clear positive correlation between the richness of information provided in the prompt and the models' performance in both error classification and explanation. Specifically, transitioning from `Zero-shot-naive` to `One-shot-detailed` leads to a marked improvement in average classification Accuracy, Macro F1 scores, and all explanation metrics. It reveals that the problem of fine-grained error analysis can derive substantial benefit from clear definitions and concrete examples. This highlights the differential impact of effective prompt engineering strategies on performance.

For the impact of Thinking Models. Comparing the performance of models with explicit "thinking" or chain-of-thought-like mechanisms, such as Claude-3.7-Sonnet-Thinking, against their base counterparts (e.g., Claude-3.7-Sonnet), reveals interesting patterns. The Thinking variant consistently achieves significantly higher Accuracy and Macro F1 scores across the different settings in the error classification task. However, their performance on error severity rating and explanation quality remains comparable to the non-thinking versions. This suggests that while structured reasoning

Table 3: Main results of the `One-shot-detailed` setting. We color the the secondary results.

Model	Think	Classification				Severity		Explanation					
		Acc↑		F_1↑		MAE↓		BLEU↑		METEOR↑		ROUGE↑	
GPT-4o	✗	66.57	72.02	52.97	57.66	0.78	0.66	2.46	2.16	19.96	19.70	28.41	26.13
o1	✓	74.49	77.80	63.94	65.57	0.82	0.69	2.27	2.61	19.04	20.49	29.42	**29.11**
o3-low	✓	**76.26**	**78.45**	**65.95**	65.20	0.77	0.69	2.30	2.01	19.96	19.04	27.79	25.61
o4-mini-low	✓	73.43	75.10	62.98	62.36	0.86	0.81	2.43	2.35	19.86	19.54	28.91	27.07
o4-mini-medium	✓	73.19	75.24	60.57	62.80	0.86	0.79	2.63	2.32	20.47	19.66	29.43	27.03
o4-mini-high	✓	73.53	74.79	64.71	62.04	0.86	0.80	2.58	2.28	20.22	19.77	28.99	27.28
Gemini-2.5	✓	76.19	77.17	65.60	64.79	0.75	0.74	**4.29**	**3.57**	**26.76**	**25.42**	**31.36**	28.06
DeepSeek-R1	✓	71.23	75.53	60.65	**66.38**	0.90	0.75	1.90	1.71	17.71	17.27	28.14	25.42
Claude-3.7	✗	69.50	75.73	55.38	62.69	**0.71**	**0.63**	3.61	3.24	22.67	22.85	29.71	27.74
Claude-3.7	✓	73.99	77.02	59.98	63.32	0.74	0.68	3.61	3.21	22.60	22.66	29.73	27.22
Grok-3	✗	67.10	74.09	52.99	60.29	0.97	0.88	3.63	2.76	24.17	23.43	29.60	26.86
Mistral-small	✗	63.64	68.68	49.69	52.93	0.94	0.81	3.06	3.33	22.14	23.53	30.50	29.26
Qwen-230b-a22b	✓	67.40	71.75	55.33	59.30	0.85	0.73	1.85	1.64	17.68	17.62	26.26	24.45
Qwen-30b-a3b	✓	63.47	68.71	47.22	54.33	1.05	0.84	1.33	1.34	16.69	17.25	26.05	24.57
Qwen-8b	✓	57.81	62.65	41.26	47.45	0.98	0.82	1.69	1.48	17.62	17.56	26.46	24.24
Llama-3.1-8b	✗	37.43	41.49	24.61	27.27	1.02	0.98	1.24	1.14	17.91	17.56	24.56	24.36
Llama-3.1-70b	✗	59.51	64.30	46.10	49.95	0.81	0.72	1.55	1.41	17.04	16.27	25.72	23.38
Llama-3.3-70b	✗	61.17	65.53	47.26	49.63	0.73	0.67	2.16	2.30	19.27	20.30	26.38	24.66
Average	-	66.85	70.61	54.22	57.13	0.86	0.77	2.41	2.21	19.95	19.83	28.10	26.16
Human	-	79.90	76.66	60.53	62.25	0.99	0.72	5.21	5.20	25.28	28.39	33.50	31.60

processes can enhance the ability to dissect and categorize errors accurately, they may not confer a similar advantage for tasks perceived as more intuitive or requiring nuanced pedagogical capability. This distinction helps isolate the specific benefits of such reasoning mechanisms and points to areas where other approaches might be needed.

For the model performance of different scale parameters. In line with general expectations from scaling laws, we observe that larger models typically yield better performance across the tasks in the `FEANEL` benchmark. For instance, within the Qwen3 series, there is a general trend of improvement as model size increases from Qwen-3-8B to Qwen-3-30B-A3B, and further to Qwen-3-230B-A22B, across all three evaluation settings. This indicates that increased capacity often leads to better generalization and task execution. However, there are exceptions to this trend. For example, Qwen-3-30B-A3B occasionally exhibits slightly superior performance on certain explanation quality metrics compared to the larger Qwen-3-230B-A22B. We attribute this to the specific pre-training and fine-tuning objectives of the Qwen-3 series, which have a strong emphasis on mathematical and coding reasoning tasks. This enhanced reasoning capability, while beneficial for many applications, may not directly or fully translate to the nuanced pedagogical communication and descriptive abilities required by our benchmark. Therefore, advancing model alignment for educational tasks and enhancing their capacity for precise, pedagogically sound error description remain significant research challenges.

3.3 Human Evaluation Results.

To gauge the performance gap between LLMs and human intelligence, we engaged several English teachers to conduct fine-grained error analysis on a randomly selected subset of 500 errors from elementary school essays and an additional 500 from secondary school essays. These teachers were not involved in the dataset construction and received no specialized training. This design choice allows us to benchmark LLM performance against the capabilities of human teachers operating without extensive task-specific instruction.

For both error classification and explanation, human teachers almost outperform all evaluated LLMs, particularly under the `Zero-shot-naive` setting. This underscores a considerable gap between

current AI capabilities and human-level expertise in the nuanced task of fine-grained error analysis. The results validate that our benchmark effectively identifies areas where LLMs have substantial room for improvement, thereby justifying its utility in driving research. While enriching prompts with detailed instructions and in-context examples does improve LLM performance and narrow this gap, LLMs still often require more extensive contextual information than humans. Moreover, their generated explanations, even when technically correct, can sometimes lack the conciseness, pedagogical appropriateness, or adaptive nuance of human-authored feedback.

3.4 Case Study

Example

Error Sentence: Mr. Cool Clothes Store's clothes is very check now.
Correct Sentence: Mr. Cool Clothes Store's clothes are very cheap now.

Ground Truth: Error Type: Subject-Verb Agreement Error, Error Severity: 4
Error Explanation: ① Subject-Verb Agreement Error: The subject ⟨clothes⟩ is plural and requires the plural verb [are], so ⟨is⟩ is changed to [are] to conform to grammatical rules; ② Adjective Error: ⟨check⟩ means to examine; the adjective [cheap] indicating price should be used here, so ⟨check⟩ is changed to [cheap] to convey the correct meaning.

GPT-4o: Error Type: Verb Choice Error, Error Severity: 4
Error Explanation: ① The verb ⟨is⟩ is incorrectly used for the plural subject ⟨clothes⟩, violating subject-verb agreement rules. The correction [are] ensures proper agreement. ② The adjective ⟨check⟩ is semantically inappropriate in this context, as it does not convey the intended meaning of being inexpensive. The correction [cheap] provides the proper adjective to convey the intended meaning.

o3-low: Error Type: Subject-Verb Agreement Error, Error Severity: 3
Error Explanation: ⟨is⟩ violates the rule that the plural subject "clothes" requires a plural verb; [are] restores subject-verb agreement. ②The word ⟨check⟩ is a misspelling/wrong form of the adjective "cheap"; replacing it with [cheap] provides the correct meaning.

Gemini-2.5-pro: Error Type: PoS Confusion Error, Error Severity: 4
Error Explanation: ① Subject-Verb Agreement Error: The subject of the sentence is "clothes," which is a plural noun. The verb used, ⟨is⟩, is singular. The verb must agree in number with its subject, so the plural verb [are] is required; ② PoS Confusion Error: The word ⟨check⟩ is used incorrectly here. The context requires an adjective following "very" to describe the clothes. ⟨check⟩ is typically a verb or a noun. The correct adjective meaning "low in price" is [cheap].

We present a case study analyzing a complex error from representative LLMs. We notice that GPT-4o misclassifies the error type for the entire edit as "Verb Choice Error." While its explanation correctly identifies the subject-verb agreement violation for "is" → "are", this is not reflected in the main error type assigned. Secondly, in its explanation for the second error component ("check" → "cheap"), GPT-4o mistakenly refers to the incorrect word "check" as an "adjective." while the semantic content of o3-mini's explanation is reasonable, the output does not fully adhere to the desired structured formatting conventions exemplified in the ground truth. For instance, it omits the explicit labeling of each error component within the explanation. Adherence to such formatting is not merely stylistic; it is crucial for consistent automated parsing of results and for providing a clear, standardized feedback experience to learners. Geimini-2.5-pro classifies the "check" → "cheap" error as a PoS Confusion Error, which is wrong since they do not have the same root or affix.

This typical case study illuminates several challenges for LLMs recurring in our experiments: the difficulty in accurately determining the single most salient and correct error type when multiple errors are present in an edit, consistent application of specific error taxonomies, and strict adherence to specified output formatting.

3.5 Effect of Prediction Order

To investigate the influence of the generation sequence on model performance, we conducted an ablation study by altering the prediction order of the three main components. Our default approach, termed *Post-explaining*, predicts elements in the sequence: Error Severity → Error Type → Error

Table 4: Ablation results of prediction order. We color the results of the secondary domain.

Setting	Pre-exp.	Classification				Severity		Explanation					
		Acc↑		F_1↑		MAE↓		BLEU↑		METEOR↑		ROUGE↑	
Zero-shot-naive	✗	61.74	66.79	46.55	52.67	0.87	0.77	**1.35**	**1.19**	17.87	**17.58**	**24.29**	**23.02**
	✓	**62.04**	**67.56**	**48.04**	**54.51**	**0.85**	**0.76**	1.31	1.07	17.97	17.32	24.37	22.99
One-shot-detailed	✗	**66.57**	72.02	**52.97**	57.66	0.78	**0.66**	2.46	**2.16**	19.96	19.70	28.41	26.13
	✓	65.90	**72.62**	51.46	**59.33**	**0.70**	**0.66**	**2.61**	2.05	**20.42**	**19.76**	**29.14**	**26.50**

Explanation. We compared this against an alternative, *Pre-explaining*, which follows the order: Error Explanation → Error Severity → Error Type. All experiments for this ablation were performed using the GPT-4o model, and the results are detailed in Table 4. In the `Zero-shot-naive` setting, the *Pre-explaining* order (generating the explanation before the error type) leads to improved performance in Error Classification and Error Severity Rating. However, this comes at the cost of reduced quality in the Error Explanation itself. This observation aligns with the intuition that outputs generated earlier in a sequence can serve as valuable contextual information for subsequent steps. Interestingly, this trend reverses in the `One-shot-detailed` setting, where the *Pre-explaining* model demonstrates nearly superior performance across all three aspects. We attribute this significant shift to the influence of the in-context demonstration. The demonstration likely provides a strong template or implicit guidance on how to effectively structure the thought process when generating an explanation first, and then coherently deriving the error type and severity from that explanation.

4 Related Work

LLMs for Education. Recent advances in LLMs have spurred a wide range of educational applications, including answer grading [46, 7], educational question generation [31, 3], interactive educational chatbots [11, 33, 53], and classroom simulation [67, 15, 64]. These systems leverage deep generative capabilities to provide personalized feedback [35, 4, 38, 66] and scaffold learners' understanding [36, 45]. For instance, LLM-based tutoring systems have shown promise in delivering real-time corrections and explanations for complex tasks [52, 28, 62]. However, ensuring that LLM-based feedback is accurate, bias-free, and pedagogically grounded remains an open challenge [63, 8]. To address these concerns, recent work focuses on pedagogical alignment [47, 42] and incorporating classical educational approaches [44, 23]. Our work fills the gap of the lack of error analysis in the context of language education.

LLMs for Language Learning. A growing body of work has specifically examined how LLMs can support second-language (L2) learners. Early efforts in automated essay scoring leveraged feature engineering or neural networks to produce holistic ratings [49, 27], while more recent studies exploit instruction-tuned LLMs to generate both scores and rubric-based rationales [6, 13]. In grammatical error correction [59], LLM prompting has been shown to narrow the gap between supervised systems and human annotators [32, 30]. Beyond sentence-level editing, researchers explore LLM-based chatbots to teach languages [41, 29]. Despite encouraging results, evaluations reveal that LLM-generated feedback may be overly generic or introduce hallucinated corrections [20, 43], underscoring the need for fine-grained and pedagogically sound analysis—a gap our benchmark explicitly targets.

5 Conclusion

This paper defines the problem of Fine-grained Error Analysis for English Learners and introduces the `FEANEL` benchmark. Our extensive evaluation of various LLMs on `FEANEL` under various prompting conditions revealed significant challenges. While current LLMs demonstrate foundational capabilities, they struggle with the consistent application of fine-grained error categories to complex student errors, often lack the pedagogical nuance and conciseness of human feedback, and exhibit performance heavily influenced by prompt engineering, model scale, and internal reasoning structures. A considerable gap to human performance persists, particularly in minimally guided scenarios, underscoring that the intricate reasoning and didactic skills essential for effective, granular educational

feedback are still developing in LLMs. We believe FEANEL serves as a crucial tool for diagnosing these limitations and will foster further research into developing more pedagogically effective and reliable AI systems for language education.

References

[1] Mistral AI. Mistral small 3.1. https://mistral.ai/news/mistral-small-3-1, 2025.

[2] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

[3] Giorgio Biancini, Alessio Ferrato, and Carla Limongelli. Multiple-choice question generation using large language models: Methodology and educator insights. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pages 584–590, 2024.

[4] Beatriz Borges, Niket Tandon, Tanja Käser, and Antoine Bosselut. Let me teach you: Pedagogical foundations of feedback for language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12082–12104, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.674. URL https://aclanthology.org/2024.emnlp-main.674/.

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

[6] SeongYeub Chu, Jong Woo Kim, Bryan Wong, and Mun Yong Yi. Rationale behind essay scores: Enhancing S-LLM‘s multi-trait essay scoring with rationale generated by LLMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5796–5814, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. URL https://aclanthology.org/2025.findings-naacl.322/.

[7] Yucheng Chu, Peng He, Hang Li, Haoyu Han, Kaiqi Yang, Yu Xue, Tingting Li, Joseph Krajcik, and Jiliang Tang. Enhancing llm-based short answer grading with retrieval-augmented generation. *arXiv preprint arXiv:2504.05276*, 2025.

[8] Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, et al. Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*, 2025.

[9] claude. Claude 3.7 sonnet and claude code. https://www.anthropic.com/news/claude-3-7-sonnet, 2025.

[10] Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. Stepwise verification and remediation of student reasoning errors with large language model tutors. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8386–8411, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.478. URL https://aclanthology.org/2024.emnlp-main.478/.

[11] Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, et al. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773*, 2023.

[12] Heejin Do, Sangwon Ryu, and Gary Lee. Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16427–16438, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.917. URL https://aclanthology.org/2024.emnlp-main.917/.

[13] Heejin Do, Sangwon Ryu, and Gary Geunbae Lee. Teach-to-reason with scoring: Self-explainable rationale-driven multi-trait essay scoring. *arXiv preprint arXiv:2502.20748*, 2025.

[14] Vacide Erdoğan. Contribution of error analysis to foreign language teaching. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, 1(2), 2005.

[15] Weibo Gao, Qi Liu, Linan Yue, Fangzhou Yao, Rui Lv, Zheng Zhang, Hao Wang, and Zhenya Huang. Agent4edu: Generating learner response data by generative agents for intelligent education systems. *arXiv preprint arXiv:2501.10332*, 2025.

[16] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[17] grok. Grok 3 beta — the age of reasoning agents. https://x.ai/news/grok-3, 2025.

[18] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[19] Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, et al. Fabric: Automated scoring and feedback generation for essays. *arXiv preprint arXiv:2310.05191*, 2023.

[20] Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student-LLM interaction. In Sachin Kumar, Vidhisha Balachandran, Chan Young Park, Weijia Shi, Shirley Anugrah Hayati, Yulia Tsvetkov, Noah Smith, Hannaneh Hajishirzi, Dongyeop Kang, and David Jurgens, editors, *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 284–293, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.customnlp4u-1.21. URL https://aclanthology.org/2024.customnlp4u-1.21/.

[21] Tao He, Hao Li, Jingchang Chen, Runxuan Liu, Yixin Cao, Lizi Liao, Zihao Zheng, Zheng Chu, Jiafeng Liang, Ming Liu, et al. A survey on complex reasoning of large language models through the lens of self-evolution, 2025.

[22] Haojing Huang, Jingheng Ye, Qingyu Zhou, Yinghui Li, Yangning Li, Feng Zhou, and Hai-Tao Zheng. A frustratingly easy plug-and-play detection-and-reasoning module for Chinese spelling check. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11514–11525, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.771. URL https://aclanthology.org/2023.findings-emnlp.771/.

[23] Thomas Huber and Christina Niklaus. LLMs meet bloom's taxonomy: A cognitive view on large language model evaluations. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5211–5246, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.350/.

[24] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[25] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[26] Carl James. *Errors in language learning and use: Exploring error analysis*. Routledge, 2013.

[27] Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. In *IJCAI*, volume 19, pages 6300–6308, 2019.

[28] Juyeon Kim, Jeongeun Lee, YoonHo Chang, CHANYEOL CHOI, Jun-Seong Kim, and Jy yong Sohn. Re-ex: Revising after explanation reduces the factual errors in LLM responses. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024. URL `https://openreview.net/forum?id=tyEWrLVU1b`.

[29] Minsol Kim, Aliea L Nallbani, and Abby Rayne Stovall. Exploring llm-based chatbot for language learning and cultivation of growth mindset. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–5, 2024.

[30] Zixiao Kong, Xianquan Wang, Shuanghong Shen, Keyu Zhu, Huibo Xu, and Yu Su. Scholargec: Enhancing controllability of large language model for chinese academic grammatical error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24339–24347, 2025.

[31] Kunze Li and Yu Zhang. Planning first, question second: An LLM-guided method for controllable question generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4715–4729, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.280. URL `https://aclanthology.org/2024.findings-acl.280/`.

[32] Wei Li and Houfeng Wang. Detection-correction structure via general language model for grammatical error correction. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1748–1763, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.96. URL `https://aclanthology.org/2024.acl-long.96/`.

[33] Anna Lieb and Toshali Goel. Student interaction with newtbot: An llm-as-tutor chatbot for secondary physics education. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2024.

[34] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[35] Anastasiya A Lipnevich and Ernesto Panadero. A review of feedback models and theories: Descriptions, definitions, and conclusions. In *Frontiers in Education*, volume 6, page 720195. Frontiers Media SA, 2021.

[36] Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. Socraticlm: Exploring socratic personalized teaching with large language models. *Advances in Neural Information Processing Systems*, 37:85693–85721, 2024.

[37] Zhexiong Liu, Diane Litman, Elaine Wang, Tianwen Li, Mason Gobat, Lindsay Clare Matsumura, and Richard Correnti. erevise+ rf: A writing evaluation system for assessing student essay revisions and providing formative feedback. *arXiv preprint arXiv:2501.00715*, 2025.

[38] Inderjeet Jayakumar Nair, Jiaye Tan, Xiaotian Su, Anne Gere, Xu Wang, and Lu Wang. Closing the loop: Learning to generate writing feedback via language model simulated student revisions. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16636–16657, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.928. URL `https://aclanthology.org/2024.emnlp-main.928/`.

[39] OpenAI. Introducing openai o3 and o4-mini. `https://openai.com/index/introducing-o3-and-o4-mini/`, 2025.

[40] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

[41] Jaekwon Park, Jiyoung Bae, Unggi Lee, Taekyung Ahn, Sookbun Lee, Dohee Kim, Aram Choi, Yeil Jeong, Jewoong Moon, and Hyeoncheol Kim. How to align large language models for teaching english? designing and developing llm based-chatbot for teaching english conversation in efl, findings and limitations. *arXiv preprint arXiv:2409.04987*, 2024.

[42] Mahefa Abel Razafinirina, William Germain Dimbisoa, and Thomas Mahatody. Pedagogical alignment of large language models (llm) for personalized learning: a survey, trends and challenges. *Journal of Intelligent Learning Systems and Applications*, 16(4):448–480, 2024.

[43] Sylvio Rüdian, Julia Podelo, Jakub Kužílek, and Niels Pinkwart. Feedback on feedback: Student's perceptions for feedback from teachers and few-shot llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 82–92, 2025.

[44] Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. How good are Modern LLMs in generating relevant and high-quality questions at different bloom's skill levels for Indian high school social science curriculum? In Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan, editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 1–10, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.bea-1.1/.

[45] Alexander Scarlatos, Ryan S Baker, and Andrew Lan. Exploring knowledge tracing in tutor-student dialogues using llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 249–259, 2025.

[46] Johannes Schneider, Bernd Schenk, and Christina Niklaus. Towards llm-based autograding for short textual answers. *arXiv preprint arXiv:2309.11508*, 2023.

[47] Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard Baraniuk. Pedagogical alignment of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13641–13650, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.797. URL https://aclanthology.org/2024.findings-emnlp.797/.

[48] Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. Exploring LLM prompting strategies for joint essay scoring and feedback generation. In Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan, editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.bea-1.23/.

[49] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1193. URL https://aclanthology.org/D16-1193/.

[50] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[51] Qwen Team. Qwen3: Think deeper, act faster. https://qwenlm.github.io/blog/qwen3/, 2025.

[52] Marcos V Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan Van Stigt, and Andre Martins. xTower: A multilingual LLM for explaining and correcting translation errors. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15222–15239, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.892. URL https://aclanthology.org/2024.findings-emnlp.892/.

[53] Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. Book2Dial: Generating teacher student interactions from textbooks for cost-effective development of educational chatbots. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9707–9731, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.578. URL https://aclanthology.org/2024.findings-acl.578/.

[54] Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.120. URL https://aclanthology.org/2024.naacl-long.120/.

[55] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*, 2024.

[56] Hanyi Xu, Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Philip S Yu. Large language models for education: A survey. *arXiv preprint arXiv:2405.13001*, 2024.

[57] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112, 2024.

[58] Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, et al. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. *arXiv preprint arXiv:2410.04509*, 2024.

[59] Jingheng Ye, Yinghui Li, Yangning Li, and Hai-Tao Zheng. MixEdit: Revisiting data augmentation and beyond for grammatical error correction. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10161–10175, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.681. URL https://aclanthology.org/2023.findings-emnlp.681/.

[60] Jingheng Ye, Yinghui Li, Qingyu Zhou, Yangning Li, Shirong Ma, Hai-Tao Zheng, and Ying Shen. CLEME: Debiasing multi-reference evaluation for grammatical error correction. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6174–6189, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.378. URL https://aclanthology.org/2023.emnlp-main.378/.

[61] Jingheng Ye, Zishan Xu, Yinghui Li, Xuxin Cheng, Linlin Song, Qingyu Zhou, Hai-Tao Zheng, Ying Shen, and Xin Su. Cleme2. 0: Towards more interpretable evaluation by disentangling edits for grammatical error correction. *arXiv preprint arXiv:2407.00934*, 2024.

[62] Jingheng Ye, Shang Qin, Yinghui Li, Xuxin Cheng, Libo Qin, Hai-Tao Zheng, Ying Shen, Peng Xing, Zishan Xu, Guo Cheng, et al. Excgec: A benchmark for edit-wise explainable chinese grammatical error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25678–25686, 2025.

[63] Jingheng Ye, Shen Wang, Deqing Zou, Yibo Yan, Kun Wang, Hai-Tao Zheng, Zenglin Xu, Irwin King, Philip S Yu, and Qingsong Wen. Position: Llms can be good tutors in foreign language education. *arXiv preprint arXiv:2502.05467*, 2025.

[64] Murong Yue, Wenhan Lyu, Wijdane Mifdal, Jennifer Suh, Yixuan Zhang, and Ziyu Yao. Mathvc: An llm-simulated multi-character virtual classroom for mathematics education. *arXiv preprint arXiv:2404.06711*, 2024.

[65] Min Zeng, Jiexin Kuang, Mengyang Qiu, Jayoung Song, and Jungyeul Park. Evaluating prompting strategies for grammatical error correction based on language proficiency. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6426–6430, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.569/.

[66] Mike Zhang, Amalie Pernille Dilling, Léon Gondelman, Niels Erik Ruan Lyngdorf, Euan D Lindsay, and Johannes Bjerva. Sefl: Harnessing large language model agents to improve educational feedback systems. *arXiv preprint arXiv:2502.12927*, 2025.

[67] Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. Simulating classroom education with LLM-empowered agents. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10364–10379, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.520/.

[68] Deqing Zou, Jingheng Ye, Yulu Liu, Yu Wu, Zishan Xu, Yinghui Li, Hai-Tao Zheng, Bingxu An, Zhao Wei, and Yong Xu. Revisiting classification taxonomy for grammatical errors. *arXiv preprint arXiv:2502.11890*, 2025.

Technical Appendices and Supplements

In this Appendix, we provide supplementary material that deepens the reader's understanding of FEANEL and facilitates full reproducibility. Appendix A reiterates the impact statement, outlining the broader educational and societal value of the benchmark. Appendix B details the complete error-type taxonomy, including prioritisation rules and illustrative examples that guide both annotation and automatic classification. Appendix C reports evaluation protocols and all prompts used in our experiments, ensuring that future work can replicate or extend our baselines without hidden hyper-parameters. Appendix D presents additional quantitative results and fine-grained analyses that complement the main paper's findings. Appendix E discusses remaining limitations of the dataset and methodology, clarifying the scope of conclusions and highlighting open challenges. Finally, Appendix F states our ethical considerations, covering data privacy, annotator compensation, and responsible release practices.

A Impact Statement

This work makes the first publicly available benchmark for fine-grained error analysis in K-12 English writing, offering expert annotations that couple each learner error with type, severity, and explanatory feedback. By standardizing an otherwise fragmented problem space, the dataset provides a reproducible test bed for both data-centric and algorithmic research, enabling rigorous cross-model comparison and accelerating progress on educational NLP. Its open-source release lowers the barrier for researchers and practitioners worldwide to study, diagnose, and improve language models' pedagogical abilities.

For educators and learners, the benchmark paves the way toward highly targeted, pedagogically sound feedback at scale. Models trained or evaluated on our data can move beyond binary correctness and supply actionable explanations that foster metalinguistic awareness, a key predictor of language

acquisition. Teachers may repurpose model-generated analyses to streamline grading, identify cohort-level pain points, and design adaptive curricula, while students gain immediate, interpretable insights into their writing weaknesses—benefits that are especially valuable in under-resourced classrooms.

Finally, the dataset invites broader inquiry into responsible AI for education. Its fine-grained labels facilitate auditing of model biases across error types and learner demographics, informing fairness interventions before deployment. Because essays are anonymised and age-appropriate, the resource balances openness with privacy. We envision the benchmark as a focal point for interdisciplinary collaboration among NLP researchers, psycholinguists, and educators, ultimately advancing equitable and effective language-learning technologies.

B Error Type Taxonomy

We illustrate our constructed error type taxonomy in Figure 3. We stipulate the priority of error types according to their top-to-bottom positions in Figure 3. For instance, Case Error is assigned the lowest priority, while Sentence Redundancy Error holds the highest. In particular, Punctuation Error is prioritized between Contraction Error and Determiner Error due to its ubiquity. Therefore, when encountering an edit involving Punctuation Error and Determiner Error, models should classify it as Determiner Error.

Examples of all error types in our proposed taxonomy are as follows:

(01) Case Error: Incorrect use of uppercase or lowercase letters. Example: Writing "paris" instead of "Paris."

(02) Space Error: Missing necessary spaces between words or having extra spaces. Example: Writing "tothelibrary" instead of "to the library."

(03) Spelling Error: The spelling of a word does not conform to standard norms. Note that both British and American spellings are considered correct and should not be classified as spelling errors. Example: Writing "recieve" instead of "receive"; "definately" instead of "definitely."

(04) Contraction Error: Incorrect use of word contractions. Example: Writing "isnt" instead of "isn't."

(05) Punctuation Error: Misuse or omission of punctuation marks in writing. For example, two or more independent clauses are improperly joined without correct punctuation or conjunctions, or sentence components that should be joined are separated into independent sentences. Example: Writing "He sings children's songs he is an excellent musician" instead of "He sings children's songs. He is an excellent musician."

(06) Determiner Error: Using inappropriate determiners or omitting necessary determiners, such as articles (a, an, the). Example: Writing "She has cat" instead of "She has a cat."

(07) Number Error: Using inappropriate cardinal or ordinal numbers. Example: Writing "two place" instead of "second place."

(08) Preposition Error: Using inappropriate prepositions or omitting necessary prepositions. Example: Writing "good in math" instead of "good at math."

(09) Auxiliary Error: Using inappropriate auxiliary verbs or omitting necessary auxiliary verbs (including basic and modal auxiliaries). Example: Writing "should sing well" instead of "can sing well."

(10) Adjective Error: Using inappropriate adjectives or omitting necessary adjectives, including improper use of comparative or superlative forms. Example: Writing "more taller" instead of "taller."

(11) Adverb Error: Using inappropriate adverbs or omitting necessary adverbs. Example: Writing "runs quick" instead of "runs quickly."

(12) Noun Number Error: Incorrect use of singular or plural forms of nouns, or confusion between countable and uncountable nouns. Example: Writing "many book" instead of "many books."

(13) Noun Possessive Error: Incorrect use or omission of possessive forms of nouns. Example: Writing "Johns book" instead of "John's book."

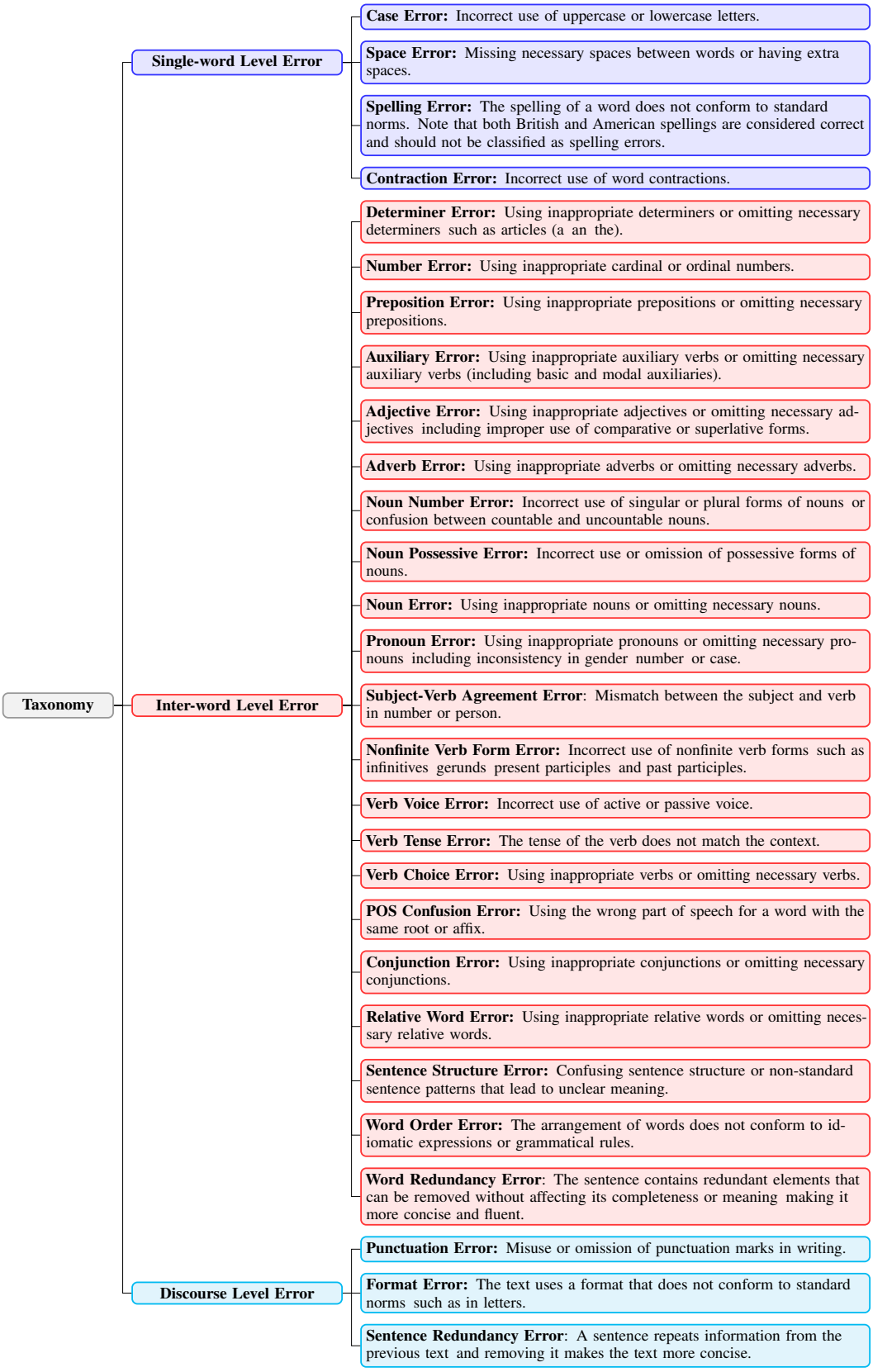


Figure 3: Taxonomy of Error Analysis for English Writing.

(14) Noun Error: Using inappropriate nouns or omitting necessary nouns. Example: Writing "The book is on the table" instead of "The book is on the shelf."

(15) Pronoun Error: Using inappropriate pronouns or omitting necessary pronouns, including inconsistency in gender, number, or case. Example: Writing "Everyone should bring their own lunch" instead of "Everyone should bring his or her own lunch."

(16) Subject-Verb Agreement Error: Mismatch between the subject and verb in number or person. Example: Writing "The list of items are" instead of "The list of items is."

(17) Nonfinite Verb Form Error: Incorrect use of nonfinite verb forms, such as infinitives, gerunds, present participles, and past participles. Example: Writing "suggested to go" instead of "suggested going."

(18) Verb Voice Error: Incorrect use of active or passive voice. Example: Writing "was ate" instead of "was eaten."

(19) Verb Tense Error: The tense of the verb does not match the context. Example: Writing "Yesterday, I go" instead of "Yesterday, I went."

(20) Verb Choice Error: Using inappropriate verbs or omitting necessary verbs. Example: Writing "tried to visit" instead of "decided to visit."

(21) PoS Confusion Error: Using the wrong part of speech for a word with the same root or affix. Example: Writing "beauty singer" instead of "beautiful singer."

(22) Conjunction Error: Using inappropriate conjunctions or omitting necessary conjunctions. Example: Writing "I wanted to go, and I was tired" instead of "I wanted to go, but I was tired."

(23) Relative Word Error: Using inappropriate relative words or omitting necessary relative words. Example: Writing "where I was born in" instead of "in which I was born."

(24) Sentence Structure Error: Confusing sentence structure or non-standard sentence patterns that lead to unclear meaning. Example: Writing "The book on the table which I read yesterday" instead of "The book which I read yesterday is on the table."

(25) Word Order Error: The arrangement of words does not conform to idiomatic expressions or grammatical rules. Example: Writing "older three years" instead of "three years older."

(26) Word Redundancy Error: The sentence contains redundant elements that can be removed without affecting its completeness or meaning, making it more concise and fluent. Example: Writing "returned back" instead of "returned."

(27) Format Error: The text uses a format that does not conform to standard norms, such as in letters. Example: Writing "Dear Sir, I am writing to you..." instead of "Dear Sir,\n I am writing to you...."

(28) Sentence Redundancy Error: A sentence repeats information from the previous text, and removing it makes the text more concise. Example: Writing "I went to the store. I went to the store to buy milk" instead of "I went to the store to buy milk."

(29) Other Error: Errors that do not fall into the above categories. Example: Non-sense sentences like "I look like beauty as famous do."

C Evaluation Details and Prompts for FEANEL

Evaluation Details. For closed-source or large models, we interact with the models through their respective APIs, ensuring consistency in input formatting and evaluation protocols. For the open-source models with parameter sizes less than 8B, we deploy them locally on NVIDIA A800 GPUs, leveraging their fine-tuned versions for conversational tasks. Each model is evaluated using identical prompts and settings to ensure fair comparisons. We set the `temperature` to 0.6 and `top_p` to 0.95.

Prompts. Our designed prompts are shown as follows:

Prompt for the `Zero-shot-naive` setting

You are an experienced English K-12 teacher, specializing in providing accurate and relevant explanations for writing errors in essays. To ensure accuracy and relevance, adhere to these principles:
1. Analysis each given edit one by one. Maintain the exact number of edits and make sure the edit index is correct.
2. Don't alter the error and correction content in any case.
3. Specify a single error type, a severity, and a description for each edit. If an edit involves multiple errors, you must predict only one error type with the highest priority order (see below). However, when describing, you must provide a detailed explanation for each error type and use numbering such as ①, ②, and semicolons to separate the descriptions of different error types.
4. Error severity is rated from 1 (trivial) to 5 (extremely serious).
5. The error description must target the predicted error type, highlight the violated semantic rules and relevant knowledge, and explain the given correction and its rationale.
6. Use specific symbols to emphasize evidence words and correction methods: (1) Evidence words from the error sentence are enclosed in ⟨⟩. (2) Correction methods from the correct sentence are enclosed in [].
7. Predict a single error type for an edit based on the following error taxonomy. Directly generate the name of the error type without serial number, e.g., "Preposition Error." Don't generate any other error types not included in the taxonomy.

Error Taxonomy:
(01) Case Error
...

8. Output strictly in the following JSON format.
{JSON Format Instruction}

Now you should deal with the following input and output a single JSON output.
{Input}

Prompt for the `One-shot-detailed` setting

You are an experienced English K-12 teacher, specializing in providing accurate and relevant explanations for writing errors in essays. To ensure accuracy and relevance, adhere to these principles:
1. Analysis each given edit one by one. Maintain the exact number of edits and make sure the edit index is correct.
2. Don't alter the error and correction content in any case.
3. Specify a single error type, a severity, and a description for each edit. If an edit involves multiple errors, you must predict only one error type with the highest priority order (see below). However, when describing, you must provide a detailed explanation for each error type and use numbering such as ①, ②, and semicolons to separate the descriptions of different error types.
4. Error severity is rated from 1 (trivial) to 5 (extremely serious).
- 1 point (trivial): Minor issues like spelling or punctuation that don't affect understanding.
Example: "I have a friand who likes football." (friand -> friend)
- 2 points (minor): Errors like verb tense or simple subject-verb disagreement that don't alter the main meaning.
Example: "He go to school every day." (go -> goes)
- 3 points (moderate): More complex errors not easy to understand, such as clause misuse.
Example: "This is the book that I told you about it." (remove it)
- 4 points (serious): Multiple issues or confusing structure that hinder understanding.
Example: "Yesterday I go store and bought some apples." (go store -> went to the store)
- 5 points (extremely serious): Errors that make the sentence incomprehensible, often due to serious word misuse or structural issues.
Example: "My brother where are playing outside cannot."
5. The error description must target the predicted error type, highlight the violated semantic rules and relevant knowledge, and explain the given correction and its rationale.
6. Use specific symbols to emphasize evidence words and correction methods: (1) Evidence words from the error sentence are enclosed in ⟨⟩. (2) Correction methods from the correct sentence are enclosed in [].
7. Predict a single error type for an edit based on the following error taxonomy. Directly generate the name of the error type without serial number, e.g., "Preposition Error." Don't generate any other error types not included in the taxonomy.

Error Taxonomy:
(01) Case Error: Incorrect use of uppercase or lowercase letters. Example: Writing "paris" instead of "Paris".
...

8. Here provides an input and output example strictly in JSON format.
{Example}

Now you should deal with the following input and output a single JSON output.
{Input}

D Extra Results

Detailed analysis of model performance on each error type. A more granular examination of model performance across individual error categories is shown in Figure 4. LLMs generally achieve high classification accuracy on frequent and structurally simple error types such as Case Error, Space Error, and Spelling Error. However, their performance significantly degrades on less frequent or long-tail categories and those requiring deeper linguistic understanding or contextual reasoning. These challenging types include Contraction Error, Number Error, Auxiliary Verb Error, Part-of-Speech (PoS) Confusion Error, Sentence Structure Error, and Format Error. This disparity underscores a key deficiency in current LLMs: an incomplete or insufficiently nuanced grasp of the full spectrum of error types defined within our comprehensive taxonomy, particularly those that are

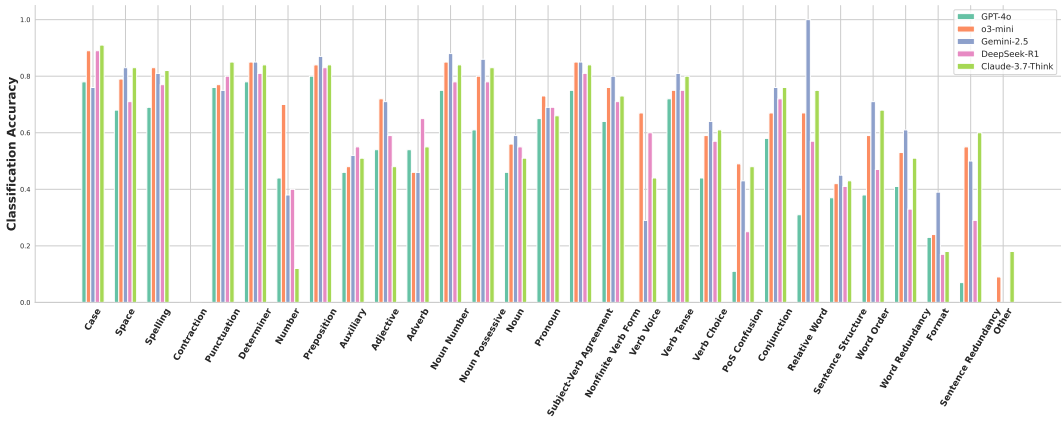


Figure 4: Classification accuracy of GPT-4o, o3-mini, Gemini-2.5-pro, DeepSeek-R1, and Claude-3.7-Think. We list the accuracy of all 29 error types.

either rare in typical training data or inherently more complex and semantically subtle, calling for future improvement.

E Limitations

While our study provides the first large-scale benchmark and systematic evaluation for fine-grained error analysis in K-12 English writing, several practical constraints and design choices limit the scope of the current work. We summarize the most salient limitations below.

K-12 focus and domain coverage. All source essays are drawn from elementary and secondary school learners. This design serves our educational goal, yet inevitably narrows linguistic variety (e.g., genre complexity, domain-specific vocabulary, discourse structures) compared with adult or professional writing. Consequently, models that perform well on FEANEL are not guaranteed to generalize to university learners, workplace communication, or other L2 populations. Extending the dataset to additional age groups, proficiency levels, and register types is a promising next step.

English-only taxonomy. Our error taxonomy is tailored to English morpho-syntax and the curricular requirements of the Chinese K-12 context. Error categories and severity rubrics may not transfer directly to other target languages or educational standards. Multilingual validation and possible language-specific extensions will be required before FEANEL can serve broader data-centric AI research in second-language learning.

Reference-based automatic metrics. We evaluate error detection, categorization, and explanation quality primarily with reference-based metrics. Although these metrics allow large-scale reproducible benchmarking, they can over-penalize legitimately different but pedagogically useful feedback, and may not fully capture fluency, readability, or learner uptake. Follow-up work should incorporate rubric-based human ratings or preference learning to complement reference matching.

F Ethics Statement

Every essay in FEANEL was scrubbed of personally identifiable information. We also ensure that no infringement or unethical behavior occurs during the dataset construction. Experienced teachers involved in the data annotation process were paid $10 - $20 per hour, which is well above local minimum wage. To maintain high-quality annotations, we developed a detailed annotation manual and conducted a pre-annotation trial, ensuring that annotators achieved at least 90% accuracy before the formal annotation process. Any annotator failing to meet this threshold was re-trained or replaced. The essay topics and texts are generally concerned with daily life. Therefore, the new research direction and tasks we propose will not cause harm to human society and education applications.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Abstract and Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix E.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Abstract, Section 3, and Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: See Abstract.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 2 and Section 3.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because running LLMs for multiple times is very expensive especially for state-of-the-art models.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 3.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: See Section F.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix A and Appendix F.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Section 2.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: See the submission page.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: See Section 3.3.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: See Appendix F.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.