LLM-as-a-tutor in EFL Writing Education: Focusing on Evaluation of Student-LLM Interaction

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim,
Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, Alice Oh
KAIST, South Korea
`{jieun_han, haneul.yoo, junho00211, 9909cindy, charlie9807, yoonsu16, takyeonlee, hwajung, juhokim, ahnsoyeon}@kaist.ac.kr`, `alice.oh@kaist.edu`

Abstract

In the context of English as a Foreign Language (EFL) writing education, LLM-as-a-tutor can assist students by providing real-time feedback on their essays. However, challenges arise in assessing LLM-as-a-tutor due to differing standards between educational and general use cases. To bridge this gap, we integrate pedagogical principles to assess student-LLM interaction. First, we explore how LLMs can function as English tutors, providing effective essay feedback tailored to students. Second, we propose three criteria to evaluate LLM-as-a-tutor specifically designed for EFL writing education, emphasizing pedagogical aspects. In this process, EFL experts evaluate the feedback from LLM-as-a-tutor regarding (1) quality and (2) characteristics. On the other hand, EFL learners assess their (3) learning outcomes from interaction with LLM-as-a-tutor. This approach lays the groundwork for developing LLMs-as-a-tutor tailored to the needs of EFL learners, advancing the effectiveness of writing education in this context.

1 Introduction

Personalized feedback is known to significantly enhance student achievement (Bloom, 1984). However, providing real-time, individualized feedback at scale in traditional classroom settings is challenging due to limited resources. Large language models (LLMs) can be particularly beneficial to address this challenge by enabling real-time feedback in educational settings (Kasneci et al., 2023; Wang and Demszky, 2023; Yan et al., 2024). However, LLMs often struggle to generate constructive feedback within educational contexts. Unlike human feedback, which consistently identifies areas for improvement, LLM-generated feedback frequently fails to effectively highlight students' weaknesses (Behzad et al., 2024). Therefore, it is essential to identify the advantages and limitations of LLMs as English writing tutors and to develop methods for providing effective feedback for students.

The evaluation of LLMs for educational purposes differs significantly from their general-purpose evaluation. General-purpose LLM evaluation primarily focuses on assessing the quality of responses (Wang et al., 2023; Zheng et al., 2023; Chang et al., 2024). However, as Lee et al. (2023) emphasize, merely evaluating the final output quality is insufficient to capture the full dynamics of human-LLM interactions. In particular, educational feedback needs a more nuanced consideration of factors beyond traditional metrics. It also requires the expertise of education professionals to evaluate the learning process and outcomes due to its inherent challenges. Our work incorporates metrics specifically tailored to pedagogical considerations by involving real-world education stakeholders to better assess student-LLM interactions.

In summary, the main contributions of this work are as follows:

1. We explore the role of LLM as tutors in generating essay feedback.
2. We introduce an educational evaluation metric customized for EFL writing education.
3. We assess student-LLM interactions by involving real-world educational stakeholders.

2 LLMs as EFL Tutors: Early Insights

In this section, we report preliminary findings that display both the advantages and limitations of LLM-as-a-tutor.

2.1 Advantage of LLM-as-a-tutor

We conduct a group interview with six EFL learners and a written interview with three instructors to explore the needs for LLM-as-a-tutor. To reflect the perspectives of key stakeholders in EFL writing education, we recruit undergraduate EFL

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
November 16, 2024 ©2024 Association for Computational Linguistics

learners and instructors from a college EFL center. The use of LLM-as-a-tutor presents a significant opportunity for EFL learners by enabling real-time feedback at scale. While all students expressed a strong need for both rubric-based scores and feedback, only two of them had previously received feedback from their instructors. Students are particularly interested in receiving immediate scores and feedback, allowing them to identify weaknesses in their essays and refine them through an iterative process.

2.2 Limitation of LLM-as-a-tutor

We conduct an experiment using `gpt-3.5-turbo` to generate essay feedback on standard setting. The model is configured to act as an English writing teacher and provide feedback based on an EFL writing scoring rubric (Cumming, 1990; Ozfidan and Mitchell, 2022). Detailed experimental settings and prompts are described in Appendix §A. We ask 21 English education experts to evaluate the feedback on a 7-point Likert scale, focusing on feedback tone (positiveness, directness) and helpfulness. The experts rate the feedback's positiveness at 5.93 and directness at 3.72. This result indicates `gpt-3.5-turbo`'s inherent tendency to generate positive feedback. However, previous research and our qualitative interviews suggest that EFL learners prefer direct and negative feedback (Ellis, 2008; Saragih et al., 2021). Moreover, the experts found the feedback from `gpt-3.5-turbo` less helpful, with an average helpfulness rating of 3.41 out of 7.

2.3 Mitigating Limitation

To address the limitations of standard prompting in generating effective feedback for EFL learners, we propose a score-based prompting method that involves informing the model of a student's essay weakness using rubric-based scores. While models like `gpt-3.5-turbo`, trained with reinforcement learning from human feedback, generally align with human preferences in broad contexts, they may not always provide the most constructive feedback for EFL learners who need more targeted guidance. These models tend to generate positive and indirect feedback, which, though satisfactory in general contexts, may not be as effective for learners who need more targeted and constructive guidance. Therefore, we suggest score-based prompting method, leveraging rubric-based scores for LLM self-refinement of feedback generation (Pan et al., 2024).

Score-based prompting method uses predicted scores and rubric explanations to generate feedback on students' essays. Student's essays are scored by the state-of-the-art automated essay scoring model (Yoo et al., 2024) under three rubrics: content, organization, and language (Table 2). We assume this scoring information can guide the model in generating feedback that is more aligned with students' needs. The exact prompting setup is described in Appendix §A.

3 Student-LLM Interaction Evaluation

In this section, we introduce evaluation methods for student-LLM interaction. We provide feedback generated with score-based prompting to student. English experts then evaluate LLM-generated feedback with our evaluation metrics on a 7 point Likert scale.

3.1 Annotator Details

We explore student-LLM interaction of 33 EFL learners and gather evaluations from 21 English education experts, who are key stakeholders in EFL writing education. These experts hold Secondary School Teacher's Certificates (Grade II) for English, licensed by the Ministry of Education, Republic of Korea. The student cohort comprises 32 Korean students and one Italian student, with a gender distribution of 12 females and 21 males. While participating in EFL writing courses, students independently write their essays, which are then subjected to LLM-generated feedback. This feedback is produced by `gpt-3.5-turbo` using score-based prompting, and is delivered through the RECIPE (Han et al., 2023) platform as part of their coursework.

3.2 Evaluation Details

We introduce educational metrics specifically designed to assess student-LLM interactions within the context of EFL writing education (Table 1). These metrics are constructed by adapting Lee et al. (2023)'s framework to fit the EFL writing settings, focusing on targets, perspectives, and criteria.

Targets We identify two primary aspects for evaluating student-LLM interactions: *output* and *process*. *Output* refers to the LLM's generated feedback that students receive, while *process* encompasses the development of students' essays, com-

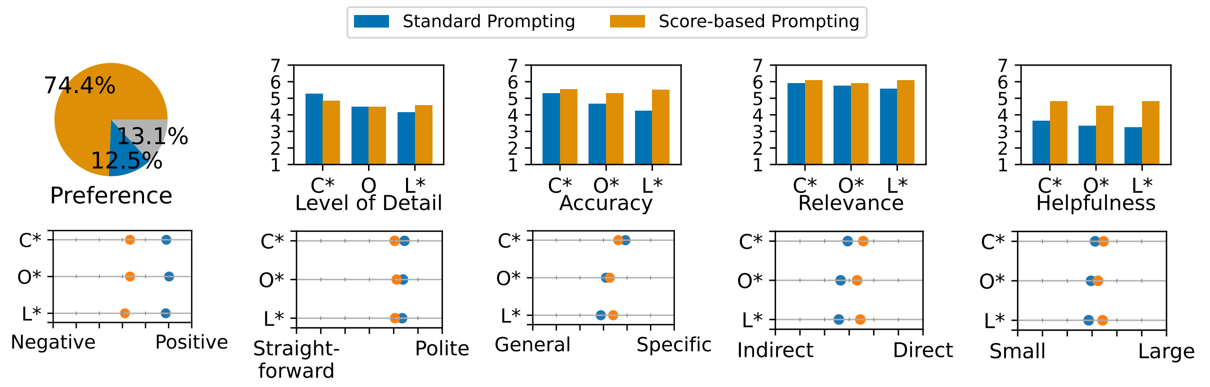

Figure 1: Evaluation results on quality and characteristic of two rubric-based feedback with standard prompting and score-based prompting in a 7-point Likert scale. C, O, and L denote Content, Organization, and Language, respectively. Asterisk denotes statistical significance tested by the paired T-test at p level of < 0.05.

Criteria	**Target**	**Perspective**	**Metric**
1. Quality	Output	Teacher	Level of detail, Accuracy, Relevance, Helpfulness
2. Characteristic	Output	Teacher	Negative-Positive, Straightforward-Polite, General-Specific, Indirect-Direct, Small-Large
3. Learning outcome	Process	Student	Essay quality improvement, Understanding

Table 1: Evaluation metrics constructed upon targets, perspectives, and criteria

prehension, and overall progress during the interaction.

Perspectives The evaluation involves the two main stakeholders in EFL education: *students* and *teachers*. While students may favor LLMs that provide immediate, correct answers, this approach may not be pedagogically optimal. Therefore, it is crucial to incorporate *teachers'* perspectives when assessing the *quality* and *characteristics* of LLM-generated feedback.

Criteria We first evaluate student-LLM interactions using three key criteria: *quality*, *characteristics*, and *learning outcomes*.

For *quality* assessment, we adapt evaluation criteria from LLM response assessments (Zheng et al., 2023), re-defining those criteria to suit our domain of feedback generation: level of detail, accuracy, relevance, and helpfulness (Appendix §B.1).

For *characteristics* assessment, we propose five characteristics to analyze the type of feedback, building on previous studies in English writing education. These criteria include: negative ↔ positive (Cheng and Zhang, 2022), straightforward ↔ polite (Lysvåg, 1975; Danescu-Niculescu-Mizil et al., 2013), general ↔ specific (Leibold and Schwarz, 2015), indirect ↔ direct (Van Beuningen et al., 2012; Eslami, 2014), small ↔ large (Liu and Brown, 2015). See Table 3 for more detailed explanations and examples. Since these five criteria are grounded in pedagogical theory and research, the analysis of feedback requires the involvement of educational experts who can interpret subtle distinctions in feedback in alignment with instructional objectives.

For *learning outcome* assessment, We assess the impact of student-LLM interaction. Students assess their own learning progress by comparing their improvement before and after receiving feedback from the LLM. After engaging with LLM-as-a-tutor to revise their essays, students reflect on their learning process through a questionnaire. The detailed questions are provided in Appendix §C.

- **Negative** ↔ **Positive**: Is the tone of feedback positive?
- **Straightforward** ↔ **Polite**: Is the feedback polite?
- **General** ↔ **Specific**: Is the feedback specific?
- **Indirect** ↔ **Direct**: Is the feedback direct?
- **Small** ↔ **Large**: How extensive is the quantity of feedback provided?

3.3 Results

In this section, we report the results of standard and score-based prompting across three criteria: *quality*, *characteristic*, and *learning outcome*.

Quality Four figures in the top row in Figure 1 present the quality evaluation results for the two types of feedback. Score-based prompting outperforms standard prompting in terms of accuracy, relevance, and helpfulness, achieving statistical significance across all rubrics. Feedback generated by standard prompting varies in the level of detail (4.16 – 5.28), while score-based prompting produces consistently detailed feedback (4.48 – 4.86). Moreover, feedback from standard prompting tends to be overly detailed in summarizing the essay, which is not perceived as constructive (see examples in Table 4). Further qualitative analysis is described in Appendix §B.2.1.

Characteristic We evaluate feedback using five metrics tailored to English writing education. Score-based prompting generates more negative, straightforward, direct, and extensive feedback compared to standard prompting across all rubrics (see the figures located in the lower section of Figure 1). Specifically, feedback from standard prompting tends to generate general compliments rather than constructive criticism. In contrast, feedback from score-based prompting is notably more concise, delivering more content in significantly fewer tokens (70.46 vs. 79.19) and sentences (4.20 vs. 5.04). To further support the results, we also conduct a qualitative analysis of the feedback characteristics on Negative ↔ Positive and Straightforward ↔ Polite (Appendix §B.2.2).

As a result, 74.38% of teacher annotators prefer feedback from score-based prompting, compared to only 12.50% who favor feedback from standard prompting (Pie chart in Figure 1). The remaining 13.12% report no difference between the two feedback types. This is statistically significant at a p level of < 0.05 using the Chi-squared test, with a fair agreement (Fleiss Kappa 0.22).

Learning Outcome The feedback provided through score-based prompting leads to a significant improvement in students' confidence regarding the quality of their essays and their understanding of each rubric (Figure 2). On average, EFL learners express high satisfaction with the LLM-generated feedback, rating 6.0 for *quality* and 6.03 for *characteristics* on a 7.0 scale. These results are statistically significant, tested by the Wilcoxon test at p value of < 0.05. Such a strong positive response underscores the potential of score-based prompting on both student confidence and satisfaction, highlighting its potential to enhance writing instruction in EFL contexts.

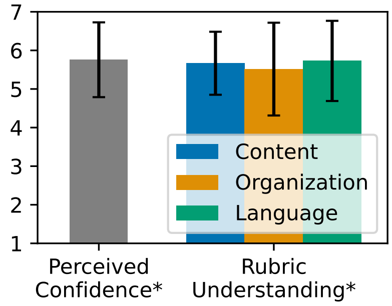

Figure 2: Learning outcome

4 Conclusion

This paper advances EFL writing education by generating and evaluating feedback tailored to students' needs, incorporating pedagogical principles, and involving real-world educational stakeholders. Our focus on essay feedback through LLM-as-a-tutor aims to more effectively support EFL students in their writing process. In the future, we plan to customize the LLM-as-a-tutor to provide individualized support. For instance, our evaluation metric and dataset can be utilized to personalize feedback, aligning with students' varying preferences. This customization would allow LLM-as-a-tutor to adapt to the specific needs and desires of each student, thereby enhancing the learning experience. Ultimately, we envision personalized LLM agents in EFL education, offering tailored support to each learner based on their unique needs.

Limitations

We utilize ChatGPT, a black-box language model, for feedback generation. This results in a lack of transparency in our system, as it does not provide explicit justifications or rationales for the generated feedback. We acknowledge the importance of and the need for continued research aimed at developing models that produce more explainable feedback, thereby opening avenues for future exploration.

Ethics Statement

We expect that this paper will make a significant contribution to the application of NLP for good, particularly in the domain of NLP-driven assistance in EFL writing education. All studies are conducted with the approval of our institutional review board (IRB). We ensured non-discrimination across all demographics, including gender and age. We set the wage per session to be above the minimum wage in the Republic of Korea in 2023 (KRW 9,260 ≈ USD 7.25) [1]. Participation in the experiment was entirely voluntary, with assurance that their choice would not influence their academic scores or grades.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government (MSIP) (No. RS-2024-00443251, Accurate and Safe Multimodal, Multilingual Personalized AI Tutors)

References

Shabnam Behzad, Omid Kashefi, and Swapna Somasundaran. 2024. Assessing online writing feedback resources: Generative AI vs. good samaritans. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1638–1644, Torino, Italia. ELRA and ICCL.

Benjamin S. Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Xiaolong Cheng and Lawrence Jun Zhang. 2022. Teachers helping efl students improve their writing through written feedback: the case of native and non-native english-speaking teachers' beliefs. *Frontiers in Psychology*, 13:804313.

Alister Cumming. 1990. Expertise in evaluating second language compositions. *Language Testing*, 7(1):31–51.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Rod Ellis. 2008. A typology of written corrective feedback types. *ELT Journal*, 63(2):97–107.

Elham Eslami. 2014. The effects of direct and indirect corrective feedback techniques on efl students' writing. *Procedia - Social and Behavioral Sciences*, 98:445–452. Proceedings of the International Conference on Current Trends in ELT.

Jieun Han, Haneul Yoo, Yoonsu Kim, Junho Myung, Minsun Kim, Hyunseung Lim, Juho Kim, Tak Yeon Lee, Hwajung Hong, So-Yeon Ahn, and Alice Oh. 2023. Recipe: How to integrate chatgpt into efl writing education. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, L@S '23, page 416–420, New York, NY, USA. Association for Computing Machinery.

Fiona Hyland and Ken Hyland. 2001. Sugaring the pill: Praise and criticism in written feedback. *Journal of Second Language Writing*, 10(3):185–212.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael S. Bernstein, and Percy Liang. 2023. Evaluating human-language model interaction. *Transactions on Machine Learning Research*.

Nancyruth Leibold and Laura Marie Schwarz. 2015. The art of giving online feedback. *Journal of Effective Teaching*, 15(1):34–46.

Qiandi Liu and Dan Brown. 2015. Methodological synthesis of research on the effectiveness of corrective feedback in l2 writing. *Journal of Second Language Writing*, 30:66–81.

Per Lysvåg. 1975. Verbs of hedging. In *Syntax and Semantics volume 4*, pages 125–154. Brill.

Burhan Ozfidan and Connie Mitchell. 2022. Assessment of students' argumentative writing: A rubric development. *Journal of Ethnic and Cultural Studies*, 9(2):121–133.

---

[1] `https://www.minimumwage.go.kr/`

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506.

Novilda Angela Saragih, Suwarsih Madya, Renol Aprico Siregar, and Willem Saragih. 2021. Written corrective feedback: Students' perception and preferences. *International Online Journal of Education and Teaching*, 8(2):676–690.

Catherine G. Van Beuningen, Nivja H. De Jong, and Folkert Kuiken. 2012. Evidence on the effectiveness of comprehensive error correction in second language writing. *Language Learning*, 62(1):1–41.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12047–12064, Singapore. Association for Computational Linguistics.

Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 626–667, Toronto, Canada. Association for Computational Linguistics.

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

Haneul Yoo, Jieun Han, So-Yeon Ahn, and Alice Oh. 2024. Dress: Dataset for rubric-based essay scoring on efl writing. *Preprint*, arXiv:2402.16733.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Appendix

A Essay Feedback Generation Model

The essay feedback generation experiments were conducted with `gpt-3.5-turbo` (0301 version) with Azure OpenAI API. To provide consistent feedback among students, we opted for a temperature setting of 0. This deterministic approach ensures that our system remains uniform, akin to evaluations from a single, consistent instructor. Below is the prompt template we used for feedback generation.

Standard Prompting

You are an English writing teacher;
give feedback on this argumentative essay with three rubrics: content, organization, and language.
`${rubric explanation}`
`${essay prompt}`
`${student’s essay}`

Score-based Prompting

You are an English writing teacher;
according to the provided score, give feedback on this argumentative essay with three rubrics: content, organization, and language.
`${rubric explanation}`
`${essay prompt}`
`${student’s essay}`

Score
`${rubric-based essay scores}`

B Essay Feedback Evaluation Details

B.1 Quality Assessment Explanation

- **Level of detail**: The feedback is specific, supported with details.
- **Accuracy**: The feedback content provides accurate information according to the essay.
- **Relevance**: The feedback is provided according to the understanding of the essay criteria.
- **Helpfulness**: The feedback is helpful for students to improve the quality of writing.

Table 2: Rubric explanations

Rubric	**Description**
Content	Paragraph is well-developed and relevant to the argument, supported with strong reasons and examples.
Organization	The argument is very effectively structured and developed, making it easy for the reader to follow the ideas and understand how the writer is building the argument. Paragraphs use coherence devices effectively while focusing on a single main idea.
Language	The writing displays sophisticated control of a wide range of vocabulary and collocations. The essay follows grammar and usage rules throughout the paper. Spelling and punctuation are correct throughout the paper.

B.2 Sample-level Analysis on Essay Feedback Evaluation

B.2.1 Quality

Table 5 shows two different language feedback examples for the same essay with a score of 2.5 out of 5.0. These examples are generated using different prompts: a standard prompt and a score-based prompt. The green text indicates detailed support and examples provided by the essay (level of detail), and the blue text describes the overall evaluation of the essay regarding the language criterion. By comparing blue text, score-based prompting suggests the improvements (helpfulness) such as *‘errors and awkward phrasing’* and *‘punctuation and capitalization’*, while standard prompting only praises language use such as *‘vocabulary and collocations’*. Considering that the language score of the essay is 2.5 out of 5.0, the feedback generated by score-based prompting appears to be more accurate. The orange text in the feedback generated by the standard prompt is irrelevant to the language criterion (relevance) and has similar expressions to an organization explanation in Table 2. We assume that score-based prompting, providing more detailed, relevant, and accurate feedback, will be more helpful to EFL students in improving their

Type	**Explanation**	**Example**
Negative	Teachers' comments indicate that there are some errors, problems, or weaknesses in students' writing.	The essay lacks depth and development in its content.
Positive	The former refers to comments affirming that students' writing has met a standard such as "good grammar", "clear organization", and "the task is well achieved".	The essay is very well-organized and effectively structured.
Polite	Politeness includes hedge expressions, modal verbs, positive lexicon, and 1st person pronouns.	However, the essay could benefit from more elaboration and development of each point.
Straightforward	Straightforward includes factuality expression and negative lexicon	The essay lacks depth and analysis.
Vague	Feedback is vague in its suggestions for ways a student can enhance their work.	There are some grammar errors.
Specific	Feedback is specific in its suggestions for ways a student can enhance their work.	There are some split infinitives in the paper. Check out more information about split infinitives in the courseroom folder titled Writing Resources.
Indirect	The teacher indicates in some way that an error exists but does not provide the correction, thus leaving it to the student to find it.	However, the essay could benefit from more examples and evidence to further strengthen the argument . . .
Direct	The teacher provides the student with the correct form.	In the third paragraph, the phrase 'unsatisfied things' could be more specific and descriptive.
Small	Feedback with a small quantity contains less content.	The essay provides a clear argument and supports it with well-developed paragraphs that are relevant to the topic. The reasons and examples provided are strong and effectively demonstrate the writer's opinion. The essay effectively addresses the prompt and provides a well-rounded argument.
Large	Feedback with a large quantity contains more extensive content in the feedback.	The essay provides a clear and well-supported argument on the topic of whether young children should spend most of their time playing or studying. The writer presents two strong reasons for their opinion that playing is better for young children. The first reason is that playing is a way of studying, as it helps children learn how to communicate and collaborate with others. The second reason is that young children are not yet mature enough for formal education, and forcing them to learn before they are ready can lead to a decline in their interest in learning. The writer supports their argument with specific examples and uses clear and concise language throughout the essay.

Table 3: Explanation and example of feedback types

essays. This again highlights the necessity of the score-based prompting process.

The green text in Table 4 represents detailed support and examples, while the blue text suggests improvements for the essay, which align with level of detail and helpfulness criteria, respectively. Feedback generated with standard prompting describes specific details of the content of the essay. However, it only summarizes the content of the essay and provides praise, instead of suggestions or criticism. On the other hand, feedback with score-based prompting points out that *'the essay lacks depth and analysis'* and *'could benefit from expanding on their ideas and providing more specific examples'*.

B.2.2 Characteristics

Table 6 presents two different content feedback examples for the same essay, scored 2.0 out of 5.0. Positive feedback is highlighted in blue text, while negative feedback is in orange. Negative feedback identifies errors in the essay, helping students recognize and address their weaknesses. Positive feedback, on the other hand, affirms that students' writing meets certain standards through praise (Cheng and Zhang, 2022; Hyland and Hyland, 2001). When comparing feedback generated by the two prompting methods, score-based prompting focuses more on negative feedback, while standard prompting only emphasizes positive aspects. Since overuse of negative feedback may discourage students' self-efficacy, teachers often precede negative feedback with positive feedback or tone down the negative feedback (Hyland and Hyland, 2001) with hedge expressions such as 'could', which is written as green text in Table 6.

C Questionaire for Learning Outcome

Please answer these questions AFTER finishing the main exercise.

1. My confidence in the quality of the essay increased after the exercise.
2. My understanding of the content criteria increased after the exercise.
3. My understanding of the organization criteria increased after the exercise.
4. My understanding of the language criteria increased after the exercise.
5. Please rate the appropriateness of the style or tone of the AI-based feedback.
6. Please rate the overall performance of AI-based scoring.
7. Please rate the overall quality of AI-based feedback.
8. Please freely share your thoughts regarding the exercise.

Standard Prompting	Score-based Prompting
The content of the essay is well-developed and relevant to the argument. The writer has identified three major problems in their country and has provided specific reasons and details to support their opinion. The writer has effectively highlighted the gender problem, inequality problem, and offensive replying culture in some communities. The writer has also provided examples of how these problems are threatening their society and splitting people up. Overall, the content of the essay is strong and persuasive.	The essay has some relevant points, but the content is not well-developed. The writer mentions three problems in Korea, but only briefly touches on each one without providing strong reasons or examples to support their argument. The essay lacks depth and analysis, and the writer could benefit from expanding on their ideas and providing more specific details.

Table 4: Example of quality evaluation on content feedback. The feedback is generated based on the student's essay, whose content score is 2.5 out of 5.0. The colors of the texts stand for quality evaluation criteria as follows: green (level of detail), and blue (helpfulness).

Standard Prompting	Score-based Prompting
The language used in the essay is sophisticated and displays a wide range of vocabulary and collocations. The writer follows grammar and usage rules throughout the paper, and spelling and punctuation are correct. The essay is well-written and easy to read. Overall, the language used in the essay is strong and effectively communicates the writer's argument.	The essay displays some control of vocabulary and grammar, but there are also several errors and awkward phrasings throughout. For example, the writer uses the phrase ('nowdays' instead of 'nowadays'), and there are several instances of awkward phrasing and word choice. Additionally, there are several errors in punctuation and capitalization. The essay would benefit from more careful proofreading and editing, as well as more attention to precise and effective language use.

Table 5: Example of quality evaluation on language feedback. The feedback is generated based on the student's essay, whose language score is 2.5 out of 5.0. The colors of the texts stand for quality evaluation criteria as follows: green (level of detail), blue (helpfulness), and orange (relevance).

Standard Prompting	Score-based Prompting
The essay effectively addresses the prompt by discussing three qualities of a good friend: honesty, sharing, and competition. Each quality is well-developed and supported with specific reasons and examples. The writer's argument is clear and persuasive, making a strong case for why these qualities are important in a friendship. Overall, the content is well-organized and relevant to the topic at hand.	The essay provides a clear and concise argument about the qualities of a good friend. The writer has identified three key qualities - honesty, sharing, and competition - and has provided specific reasons and examples to support each point. However, the essay could benefit from more elaboration and development of each point. For example, the writer could provide more specific examples of how honesty is important in a friendship, or how competition can lead to personal growth. Additionally, the essay could benefit from a stronger conclusion that summarizes the main points and provides a final thought on the topic.

Table 6: Example of characteristics analysis on language feedback. The feedback is generated based on the student's essay, whose language score is 2.0 out of 5.0. The colors of the texts stand for feedback type as follows: blue (positive), green (polite), and orange (negative).