Koh, E., et al. (Eds.) (2026). Proceedings of the 2nd International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education.

Prompt Engineering for Automated Essay Scoring in Higher Education: A Case Study in Academic Reading

Yu YAN[a*], Changhao LIANG[b], Yu-Tung CHEN[a] & Hiroaki OGATA[b]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b] *Academic Center for Computing and Media Studies, Kyoto University, Japan*
yan.yu.83x@st.kyoto-u.ac.jp

Abstract: This study proposes a prompt engineering framework leveraging GPT-4o to automate the evaluation of academic reading reports. Integrating Role-Setting, Few-Shot Calibration, and Structured JSON constraints, our method achieved a strong correlation with human grading (r = 0.83, p<0.001) and reduced Mean Absolute Error by 52.2%. The framework successfully quantified a significant "learning gain" (+5.80 points) between pre- and post-revision drafts, offering a scalable solution for process-oriented mentorship in collaborative learning.

Keywords: Large Language Models, Automated Essay Scoring, Prompt Engineering

1. Introduction

Assessing intermediate drafts in peer review is bottlenecked by manual grading workloads (Baidoo-Anu & Owusu Ansah, 2023), making it difficult to quantify "learning gains" (Gašević et al., 2015). Standard zero-shot LLMs struggle with complex rubrics (Mizumoto & Eguchi, 2023). We introduce a composite prompting strategy to evaluate drafts automatically, shifting assessment from summative judgment to process-oriented mentorship.

2. Methodology

We evaluated 81 submissions from 27 students in a Learning Analytics course. Our framework assigns GPT-4o a "Professional University Teaching Assistant" persona, enforcing contextual isolation among the source paper, student submission, and logic-based grading criteria. We embedded seven few-shot examples for calibration. A structured Chain-of-Thought (CoT) process and JSON schema constraints forced internal validation. To ensure stability, we used the median score from five repeated evaluations.

3. Evaluation and Results

We performed a correlation analysis between the scores generated by the proposed model and the ground truth scores. Figure 1 illustrates the alignment between human and AI evaluations. The analysis yields a Pearson correlation coefficient of $r = 0.83$, indicating a moderate-to-strong positive linear relationship. The statistical significance is confirmed ($p < 0.001$), rejecting the null hypothesis of independence.

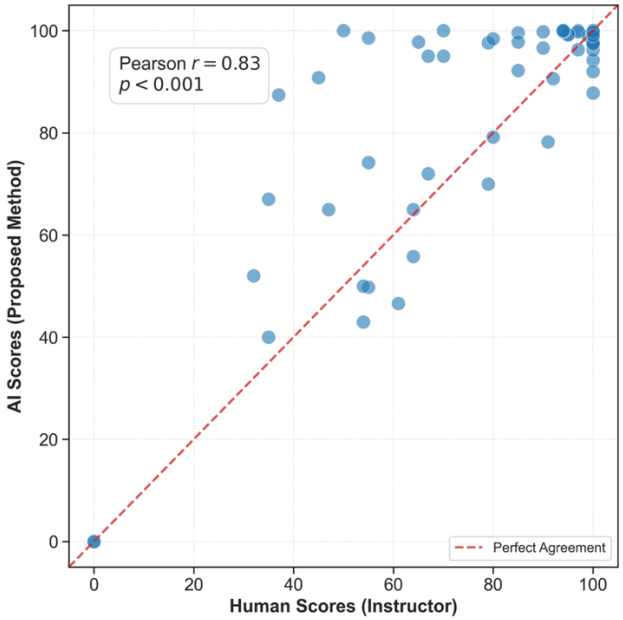

Figure 1. *Consistency Analysis (Human vs AI)*

To quantify the contribution of each prompt engineering component, we conducted an ablation study. Table 1 summarizes the performance metrics across three stages.

Table 1. *Comparative Analysis of Scoring Accuracy and Stability*

	MAE	RMSE	STD
Ablation1 (Role)	16.56	19.78	10.82
Ablation2 (Role + Few-Shot)	11.77	14.80	8.99
Proposed (Full Model)	7.92	14.48	12.12

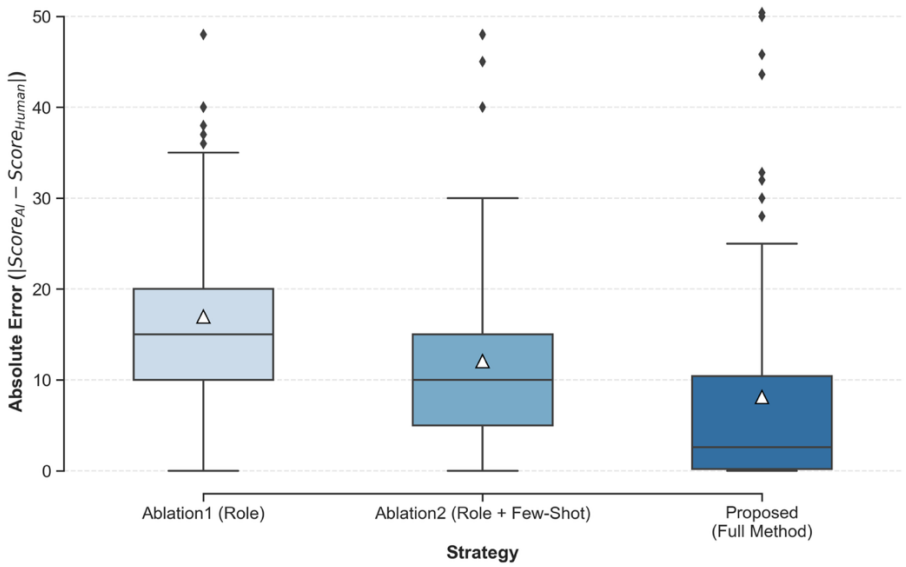

Figure 2. *Ablation Analysis of Scoring Deviation*

To operationalize the assessment of student progress, we applied the calibrated model to the complete dataset of unassessed drafts. After excluding non-submission outliers, specifically instances where students missed the initial draft but submitted the final version (or vice versa), the analysis focused on 74 valid paired observations. As shown in Figure 4, a paired sample t-test confirms a statistically significant improvement in content quality ($p <$

0.001). The mean score increased from 83.60 to 89.40, demonstrating a measurable net learning gain of 5.80 points.

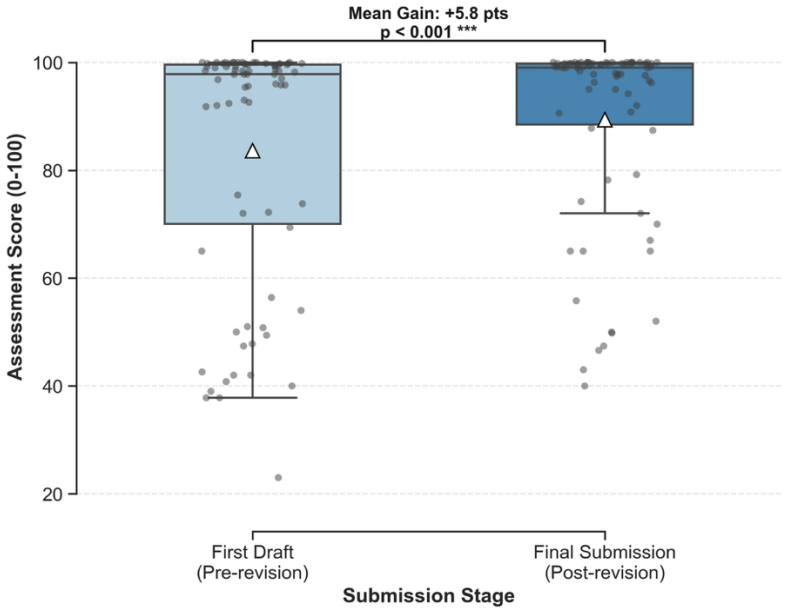

Figure 3. *Comparison of AI-generated scores between Initial Drafts and Final Submissions*

4. Discussion and Conclusion

This study validates an automated assessment framework capable of overcoming the evaluation bottleneck inherent in collaborative learning. By enforcing "structural decoupling" through JSON constraints, our methodology shifts the LLM's processing focus from creative writing to rigorous logic verification. This effectively curbs the "fluency bias" of generative models, enabling GPT-4o to internalize complex rubrics and achieve expert-level reliability. Removing the prohibitive manual workload of grading intermediate drafts facilitates a critical pedagogical shift: moving from summative judgment to process-oriented mentorship. As demonstrated, the model successfully operationalized the measurement of learning gains, confirming its sensitivity to genuine content improvements.

However, a qualitative inspection reveals a trade-off between rubric adherence and pedagogical intuition. The LLM's strict "evidence-first" logic occasionally penalizes "implicit argumentation". While this minimizes hallucination, the model may underestimate intellectually profound but structurally informal critiques that human graders naturally reward. Building on this validated framework, future work will correlate AI-measured learning gains with specific textual peer feedback to uncover which types of comments most effectively drive improvements in academic literacy.

Acknowledgements

This work was supported by the Council for Science, 3rd SIP (JPJ012347) and JSPS KAKENHI Grant Number 25K21357.

References

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52-62.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, *59*(1), 64-71.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050.