INTERNATIONAL JOURNAL OF TEACHER EDUCATION RESEARCH STUDIES (IJTERS)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3049-1614



An Automated Exam Marking Framework Using an Artificial Intelligence Tool for English Writing in General Foundation Program

Nauman Ahmad

Department of Mathematics and I.T, Centre for Preparatory Studies, Sultan Qaboos University, Sultanate of Oman

Article Information

Received: 3rd September 2026
Received in revised form: 15th September 2026
Accepted: 5th October 2026
Available online: 18th October 2026

Volume: 3
Issue: 6
DOI: https://doi.org/10.63090/IJTERS/3049.1614.0047

Abstract

This study proposes a fully automated exam marking framework for English writing in the General Foundation Program (GFP) through the use of prompt engineering with artificial intelligence (AI) tool ChatGPT and computer-based analysis. The AI tool generated 120 GFP B2 level writing responses (120 to 150 words each) with intentional grammatical errors. Four prompts were applied for automated marking by the AI tool as an examiner, based on a provided rubric, to detect errors, and awarding of a score [/10] (out of 10) to each response. A combination of different prompting techniques, including instruction-based, role-based, and formatting prompting were collectively used to design each prompt. All of the 120 responses were marked by each prompt. The first and second prompts were similar, just with a difference of paraphrasing. Prompts 3 and 4 were based on Prompt 2, containing only an addition of a single word for Prompt 3, and replacement of the added word with another word for Prompt 4. Statistical comparisons were done in SPSS. The Friedman test showed the similarities (Prompts 1 and 2 with paraphrased versions) and significant differences (Prompts 2, 3 and 4; $p < .001$, just due to an addition, and replacement of a single word) among the prompt's outcomes based on the markings. Pairwise comparisons were performed using the Bonferroni correction to check for significant differences, while Kendall's W test was used to measure the strength of agreement among the rankings. Human involvement in the study was limited to author-led validation of all the 120 responses, toward the accuracy of intentional added errors. The framework demonstrated how different prompts impacted the marking outcomes, which highlighted the sensitivity of the prompts to the addition of a single determinative word and the replacement of that word with another. Moreover, statistical comparisons showed how scalable assessment design and quality checks can be consistently applied for GFP English writing. This study showed that the proposed automated marking framework can support consistent, scalable, and prompt-sensitive assessment of GFP English writing.

Keywords:- Automated Grading, ChatGPT, Foundation Program, English Writing, Prompt Engineering, Prompt Sensitivity.

I. INTRODUCTION

In the advent of artificial intelligence, education sector is using these tools to get the help in teaching, learning and assessment. ChatGPT as a generative AI tool has a strong ability to understand, create, and evaluate responses in common language. There is an increasing interest in using AI tools for written assessment because these tools can make the process of marking faster, easier to manage, and more consistent.

In English language teaching and learning, assessment of writing is still a difficult task. The process of marking takes a lot of time and depends greatly on the judgments of human raters. Scores can be affected because of raters different interpretation of the rubric, fatigue, or difference in their experience that can result in scoring inconsistency. These challenges are especially common in large English language programs such as the General Foundation Program, where many student writing tasks are required to be evaluated within a short period of time. Consequently, there is growing interest in automated marking systems that can support assessment practices, maintain fairness, consistency, scalability, and reliability.

The effectiveness of AI-based assessment depends on the ability of the AI tool, and how clearly the instructions or prompts are written and given to it. This process is known as prompt engineering that is very important in shaping AI-generated outcomes and evaluation results. Moreover, small changes in the wording of a prompt can affect how an AI tool recognizes

errors, understands a rubric, and gives scores. This sensitivity raises important concerns about the reliability and consistency of AI-assisted marking, especially when these tools are used in educational assessment settings.

Many research studies have worked on AI-based writing assessment, but only a few studies have looked at how small changes in prompts can have an impact. Most research mainly discusses how AI tools can be used in assessment; however, there is less focus on testing and evaluating different prompt versions on the same pieces of writing. This gap is important because consistency is necessary in every assessment system, whether the marking is done by humans or by AI tools.

Therefore, this study proposes a novel automated marking framework and explores the effect of small prompt variations on automated marking of English writing responses in the GFP setting to support scalability, consistency, and reliability. The research uses different prompt versions, including paraphrased version, word addition, and word replacement. It examines how small changes in designing prompts can affect grammatical error detection and scoring. Statistical analysis is used to determine whether the differences in marking outcomes of different prompts are significant. By focusing on prompt sensitivity, the study aims to contribute to the development of more reliable, transparent, and scalable AI-supported approaches to support the process of writing assessment.

The objectives of this study are as following:

- To understand how different AI-based prompt designs with minor variations affect the grading of writing assessment.
- To compare the marking results of writing responses generated by different prompts and check if the differences are statistically significant.

II. LITERATURE REVIEW

2.1. General Foundation Program (GFP) in Oman

The General Foundation Program (GFP) is a preparatory program in Oman designed to develop students' essential skills before they move to their higher studies (University of Technology and Applied Sciences, n.d.). It focuses on areas of English language, mathematics, computing, and study skills.

2.2. What is Artificial Intelligence?

Artificial intelligence is a field of computer science that creates intelligent programs or machines capable of doing human-like tasks, such as problem-solving and decision-making, using algorithms and statistical models to analyze and interpret data (Layton, 2024).

2.3. Artificial Intelligence and Fifth Industrial Revolution

There is a close connection between Artificial Intelligence and Fifth Industrial Revolution. AI is a system of computer programs applied on machines that imitate or pretend like human intelligence. The computer programs are centered on step-by-step instructions, called algorithms that allow them to do tasks and decision-making (Sharma & Bhargav, 2022). AI has an amazing capability to do transformative alteration, and play a vital role in fifth industrial revolution and refining all areas of organizational processes (Kassir, 2023). Artificial intelligence is going to play an important role in moving and helping the progressions of Fifth Industrial Revolution (Göcke et al., 2020).

2.4. What is Generative AI?

Generative AI refers to a type of artificial intelligence that uses deep learning algorithms to generate new data, such as text, images, video or music that mimics human-created content. It is capable of creating original content without the need for human input, and is used in applications such as chatbots, content creation, and even art (Dhamani and Engler, 2024).

2.5. Prompt, Prompt Engineering, and Chat GPT

A prompt is a question or instruction that is given to an Artificially Intelligent Program in the form of text to get an answer or a response (Stirling, 2024). Prompt Engineering means asking clear and specific questions to help AI program in giving accurate and helpful answer according to our requirements (Stirling, 2024). ChatGPT is an AI interface in which a user can write questions or instructions in natural language to generate responses. In education it can be used to respond to various topics by answering user's questions and generating text (Moorthi et al., 2023).

2.6. Types of Prompts

2.6.1. Korzynski et al. (2023) defined main elements of designing effective prompts as following:

- Instruction-based Prompting- When direct instruction is given to the AI program on what to do.
- Contextual Prompting- When background or context is provided to guide the AI program.
- Formatting Prompting- When a desired format of the response is requested from the AI program.

2.6.2. Role-based Prompting

When an AI program is requested to take on a specific role to generate the outcomes; it can be done by defining the role, defining task and defining constraints (Chen et al., 2025).

- Chain-of-Thought Prompting- When AI is encouraged to think step-by-step to solve complex problems (Stirling, 2024).
- Exploration Prompting- When AI program is requested to brainstorm or generate creative ideas (Arenes, 2024).

2.7. AI in Education and Language Assessment

Today, AI is an important part of our daily routine, and has changed the way we live (Cantú-Ortiz et al., 2020; Dai et al., 2020). AI literacy, usage and skill development is extremely important today, and people those have the skills to use AI will have an edge over those without AI skills ; therefore, inclusion of AI in education is important (Cantú-Ortiz et al., 2020; Ng et al., 2021).

In the field of education, it is widely known that automated grading frameworks emphasize on the importance of validity provided by evaluating system, its reliability, and overall performance in assessment contexts, similar to manual grading system. Major factors of automated grading system are alignment with assessment goals, consistency in grading, and the ability to produce stable results.

Chapelle and Voss (2016) stated that automated writing evaluation (AWE) that is also known as automated essay evaluation (AEE), is a technology which is designed to make the process of assessing students' English writing faster and more efficient. The development of these systems has been a complex area of research that brings together different fields within language assessment.

The use of artificial intelligence tools for language assessment is gaining popularity, especially in automated essay evaluation (AEE) and automated writing evaluation (AWE). AI tools help in evaluation of written texts, make the process faster and provide quick feedback to learners (Shermis & Burstein, 2013). Moreover, research has demonstrated that automated assessment systems can reduce time and cost, decrease the workload of human raters and maintain acceptable levels of marking consistency (Page, 2003; Dikli, 2006).

Shermis and Burstein (2013) indicated that the software like e-rater is used for automated essay scoring, which can evaluate written text through the detection and examination of grammar, coherence, and vocabulary with a strong level of consistency. Guo et al. (2024) highlighted the increasing role of AI-powered systems in large scale assessment through their support towards improvement in efficiency and reduction in cost and time as compared to traditional evaluation methods. In addition, studies by Chen et al. (2024) mentioned that AI driven tools are moving toward multimodal assessment that supports the language skills assessment through a combination of text, audio, and visual inputs.

Susilo and Firdaus (2025) stated that AI tools are supporting the process of assessment through automation, personalization, and multimodal evaluation. These developments in AI automation improve the overall efficiency, and as a result the evaluation of large-scale assessments become easy that can be maintained with established grading standards. Consequently, today automated systems are in high demand, and used increasingly in language evaluation settings. However, there are still many challenges that need to be addressed, including variation in automated scoring, and limitation in transparency; thus, they are affecting the consistency and reliability of automated scoring systems. There should be AI-based auto scoring systems those could be implemented in an organized manner to meet these challenges. Future research should focus on reliable AI-based evaluations for linguistic and contextual settings that supports reliable and scalable language assessment.

Writing evaluation is complex in English language assessment because of its subjective nature. Human marking is often different because of different raters interpretations of marking rubrics that results in marking variability. Studies on writing assessment have shown that inter-rater reliability is a continuous challenge, which is influenced by examiner experience, fatigue, and contextual factors (Weigle, 2002).

Yang et al. (2024) explained that Automatic Essay Scoring (AES) uses machine learning to assess students' writing. Most studies focus on models made for one specific prompt or on models that work across different prompts. Their study was based on 25,000 essays, and found that prompt-specific models give more accurate results but may create more bias. However, cross-prompt models work better across different tasks. The findings showed that AES performance is prompt dependent that raises concerns about consistency, fairness, and reliability in automated assessment systems.

Liew and Tan (2025) explored in automated essay grading (AEG) through the use of large language models (LLM), including GPT-3.5 and GPT-4 that was based on automated scoring and feedback provision. This study indicated that LLM-based systems can achieve significant agreement with human raters. Consequently, they have strong potential for improving efficiency and consistency in writing assessment. These findings highlighted the effectiveness of AI-based grading. However, they provided limited insight towards the consistency of these systems under varying prompt conditions. This raises an important question regarding the reliability of AI-generated scores, especially when prompt design is changed.

Existing studies show that making automated essay scoring (AES) systems fair and reliable is a major challenge. There are many models trying to improve accuracy, but their outcomes may be biased. In addition, studies on large language models (LLMs) show that their scoring does not always match human raters and can change depending on the type of essay. For example, LLMs can give higher scores to shorter essay and lower score to longer essay, which proves that their evaluation methods are different from human judgement. Though LLMs provide consistent feedback and scoring patterns, these findings raise concerns about fairness, consistency, and agreement with human judgement. These issues can also be affected because of prompts design in AI-based assessment (Mathew et al., 2026; Schaller et al., 2024). Therefore, it is important to carefully evaluate and design these AI-assisted grading systems before using them in educational environment.

Huang et al. (2026) examined the accuracy and fairness of generative AI-based automated essay scoring through different prompting strategies and found that the prompts those were designed carefully could significantly provide better results and elevated alignment with human ratings. The study showed that AI models can reach human-level scoring performance; however, prompt variation and model configuration affect accuracy and fairness. These findings highlighted the importance of designing prompts to influence grading outcomes and raise concerns about consistency and reliability in AI-based assessment. However, exiting research has not sufficiently examined how micro-level prompts variations, like change of a single-word, replacement of a word with another, or minor paraphrasing may influence automated scoring outcomes, which remains an important gap that is addressed in this study.

Salinas and Morstatter (2024) explained that the wording of prompts can significantly change the outcomes of LLM model. Researchers tested different prompt variants and found that even a very small change like adding an extra space in a

prompt can change the response of LLM. The study emphasized that prompt design strongly effects LLM's reliability and consistency. Therefore, prompt engineering should be done with the utmost care, keeping in view the impact of even small changes.

There are many studies on prompt engineering, but only a few of them have examined how small changes in prompts can affect automated grading results. Most studies focused on the improvement of the quality of outcomes instead of focusing on consistency and care towards different prompt versions. This is an important gap, especially in educational assessment, where reliability and consistency are very important.

Therefore, there is a need for systematic investigation to explore how small prompt variations can influence AI-based assessment. Filling this gap can contribute towards the improvement of reliability and trustworthiness of automated assessment systems. This study responds to this need by examining the effects of prompt paraphrasing, word addition, and word replacement on automated marking outcomes in English writing assessment within the General Foundation Program setting.

III. METHODOLOGY

The study focused on developing an automated marking framework that applied different prompt designs in ChatGPT to assess GFP English writing responses. The results were statistically analysed using IBM SPSS Statistics to measure the impact of quantitative experimental prompt variations on automated outcomes of language grading. There were 120 writing responses generated through ChatGPT to keep the evaluation process consistent and automated, with each response containing 120 to 150 words. These responses were designed to reflect General Foundation Program B2 level writing that included intentional grammatical errors to represent genuine performance by students.

There were four prompts used for the study. These prompts mainly used instruction-based prompting, where clear steps were given to the AI-tool. In addition, they also included role-based prompting by asking the AI-tool to act as an examiner, and formatting prompting to keep the output in the form of graded marks; including, Prompt 1). Base prompt for marking the writing responses as an examiner, keeping in view a provided rubric for grading the written text. Prompt 2). Paraphrased version of Prompt 1. Prompt 3). Slightly changed version of Prompt 2 (addition of a new word 'strictly'). Prompt 4). Minor replacement in Prompt 3 (newly added word 'strictly' was replaced with the word 'leniently'). The prompts are provided in the Appendix A.

Each writing response was evaluated using all four prompts. The AI tool ChatGPT detected grammatical errors (tense, subject–verb agreement, article misuse, punctuation, word order), and assigned a grammar score out of 10 marks on the basis of rubric (10 marks= Perfect grammar. 8 to 9 marks = Small mistakes, meaning still clear. 5 to 7 marks = Several mistakes, clarity sometimes reduced. 2 to 4 marks = Many major mistakes, hard to read. 0 to 1 mark = Grammar is very poor and unclear).

The researcher reviewed all the writing samples for full validation of error detection accuracy. The Friedman test was used that is a non-parametric statistical test to compare the marking outcomes across the four prompts, to conclude whether differences were statistically significant. Pairwise comparisons were done using the Bonferroni correction to measure the significance, and Kendall's W test was used to gauge the effect size and level of agreement among the rankings.

Following are the research hypotheses:

- The null hypothesis (H_0) states that there is no significant difference in the grades of writing assessment given by different prompts.
- The alternative hypothesis (H_1) states that there is a significant difference in the grades of writing assessment given by different prompts.

IV. RESULTS

To test the significant differences and performance among four prompts (Prompt 1, Prompt 2, Prompt 3, and Prompt 4), results of 480 auto graded writing responses were compared using the Friedman test in SPSS. The SPSS results showed a significant difference between the prompts, $\chi^2(3) = 351.23$, $p < .001$. Table 1 and Fig 1. show the mean ranks of prompts. Prompt 4 has the highest mean rank (M = 3.99) where the word 'leniently' was used, followed by Prompt 1 (M = 2.49) and then Prompt 2 (M = 2.46), those were paraphrased versions. Prompt 3 has the lowest mean rank (M = 1.06) where the word 'strictly' was used. Further, pairwise comparisons using the Bonferroni correction showed that all three differences were significant ($p < .001$), except for Prompt1 and Prompt 2 ($p = 1.000$) that performed similarly. Table 2 shows pairwise comparisons results for hypotheses testing. The Kendall's coefficient of concordance showed an extremely high level of agreement among the rankings, $W = .976$, which indicates near-perfect consistency in the evaluation outcomes across four prompts, as shown in Table 3.

Table 1. Mean Ranks of Prompts

Prompt	Mean Rank	Prompt Variant
Prompt 4	3.99	Replacement of the word 'strictly' with the word 'leniently'
Prompt 1	2.49	Base Prompt
Prompt 2	2.46	Paraphrased version of Prompt1
Prompt 3	1.06	Addition of the word 'strictly'

Fig 1: Mean Ranks of Prompts

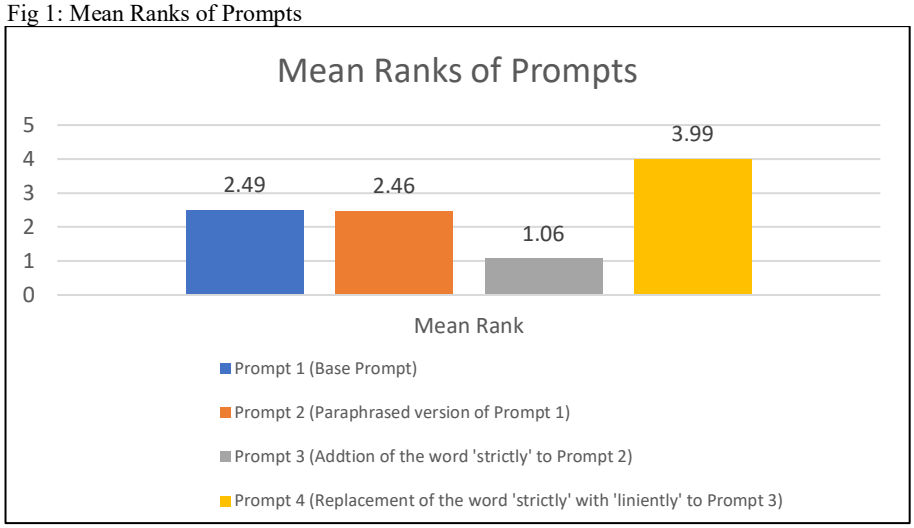


Table 2. Pairwise Comparison Results of Hypothesis Testing

Comparison	Significant	Decision
Prompt 1 vs Prompt 2	No ($p \geq .001$)	Accept H_0 (no difference)
Prompt 1 vs Prompt 3	Yes ($p < .001$)	Reject H_0 (different)
Prompt 1 vs Prompt 4	Yes ($p < .001$)	Reject H_0 (different)
Prompt 2 vs Prompt 3	Yes ($p < .001$)	Reject H_0 (different)
Prompt 2 vs Prompt 4	Yes ($p < .001$)	Reject H_0 (different)
Prompt 3 vs Prompt 4	Yes ($p < .001$)	Reject H_0 (different)

Table 3. Kendall's W Test (Coefficient of Concordance)

Kendall's coefficient of concordance	**Value**
	W = .976

V. DISCUSSION

The results of this study clearly indicate that prompt designing play a significant role in shaping the automatic AI-generated marking outcomes. It was evident that a minor addition of a word ('strictly') to the prompt, and replacement of the newly added word ('strictly' with 'leniently') changed the grades for each prompt category to measureable differences; except Prompts 1 and 2, as they were paraphrased versions of each other. This outcome aligns with the existing research on prompt engineering in large language models. The findings clearly reveal that a small lexical change in prompt wording can impact the outcome of the AI tool, interprets the marking rubric differently, and assigns different scores to writing responses. In this study, Prompt 3 produced the lowest mean rank; however, Prompt 4 produced the highest mean rank. The prompts differ only in the use of the words 'strictly' and 'leniently'. This outcome clearly suggests that AI-generated grading outcomes are highly sensitive to instructional language of the prompt.

The Friedman test results showed statistically significant differences; $\chi 2$ (3) = 351.23, $p < .001$, and resulted to the rejection of the null hypothesis for prompts 2, 3 and 4; pairwise comparisons revealed that three prompt variations produced significantly different results except the comparison between Prompt 1 and Prompt 2, which verify that paraphrasing did not significantly impact scoring performance. However, the addition and replacement of a single evaluative word produced different scoring results. These findings emphasize that prompt structure and wording should be carefully planned for AI-based assessments.

Moreover, the research highlights the potential of prompt engineering that can optimize the outcomes of automated marking frameworks. It explains that if prompts are carefully planned and refined, teachers and evaluators can improve consistency and reliability that will make AI-based assessment more feasible and practical for large-scale assessments. The very high Kendall's coefficient of concordance (W = .976) proved a near-perfect level of agreement among the rankings of prompts. This result specifies that even if score levels changed between prompts, the overall ranking pattern of responses was highly consistent. These findings highlight the contribution of the framework, and recommends that AI-based marking systems can be used for stable evaluation performance keeping in view the sensitivity of prompt wording. Therefore, prompt engineering should be done with extreme care to maintain fairness, reliability, and consistency for AI-based writing assessment, especially if it is prepared for large-scale educational settings like the General Foundation Program.

VI. LIMITATIONS

This study has some limitations. The writing responses were AI-generated and tested in a controlled setting using ChatGPT. In addition, no human raters were included for direct comparison with the automated scores. However, the controlled design helped in clearly examining the effect of prompt variations on automated grading outcomes.

VII. CONCLUSION

This study validates that a fully automated marking framework using ChatGPT can effectively evaluate and grade GFP-level writing; however, its performance is highly dependent on the design of prompt. The outcomes of the research emphasize that it is very important to carefully design the prompts to ensure consistency and fairness in AI-based assessment. The findings of the Friedman test confirmed that even small variations in prompt wording can significantly affect grading outcomes that highlights the sensitivity of prompt designing in AI-based assessment systems.

Moreover, the high Kendall's coefficient of concordance also verified a near-perfect level of agreement among prompt rankings, and suggested that AI-assisted marking can maintain stable evaluation patterns, if the prompts sensitivity is carefully considered. This research is an empirical contribution to the growing body of literature towards the area of AI in education; linked to prompt sensitivity and prompt engineering. It provides practical insights for designing scalable and reliable AI-based automated marking systems. Overall, the findings suggest that the proposed automated marking framework can provide a consistent and scalable way to evaluate GFP English writing while emphasizing how small prompt changes can impact marking outcomes.

Future research should include and explore real data of students, and make comparison with human raters. This framework can be expanded toward writing tasks other than grammar. Moreover, Contextual Prompting, Chain-of-Thought Prompting, and Exploration Prompting can also be used for the future studies, to explore other types of prompts. Further studies may also examine the effect of prompt variations toward different proficiency levels and educational settings to understand the reliability and fairness of AI-based evaluations.

DECLARATIONS

Ethical Statement

Ethical approval was not required for this study because no human participants or personal data were involved. The AI-generated writing responses and automated evaluations were reviewed and validated by the author in a controlled experimental setting.

REFERENCES

Arenes, S. (2024). *Prompt engineering: Advanced strategies for optimizing language models like ChatGPT*. SA Publisher.

Cantú-Ortiz, F. J., Galeano Sánchez, N., Garrido, L., Terashima-Marin, H., & Brena, R. F. (2020). An artificial intelligence educational strategy for the digital transformation. *International Journal on Interactive Design and Manufacturing (IJIDeM), 14*(4), 1195–1209. https://doi.org/10.1007/s12008-020-00702-8

Chapelle, C. A., & Voss, E. (2016). 20 years of technology and language assessment in *Language Learning & Technology, 20*(2), 116–128. http://llt.msu.edu/issues/june2016/chapellevoss.pdf

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns, 6*(6), Article 101260. https://doi.org/10.1016/j.patter.2025.101260

Chen, C., Tang, M., & Yang, A. (2024). Progress in a multimodal learning evaluation based on artificial intelligence. *Frontiers in Educational Research, 7*(7). https://doi.org/10.25236/fer.2024.070704

Dai, Y., Chai, C. S., Lin, P. Y., Jong, M. S. Y., Guo, Y., & Qin, J. (2020). Promoting students' well-being by developing their readiness for the artificial intelligence age. *Sustainability, 12*(16), 6597. https://doi.org/10.3390/su12166597

Dhamani, N., & Engler, M. (2024). *Introduction to generative AI*. Manning Publications.

Dikli, S. (2006). Automated essay scoring. *Turkish Online Journal of Distance Education, 7*(1), 49–62.

Göcke, L., Soltanifar, M., & Hughes, M. (2020). *Digital entrepreneurship: Impact on business and society*. Springer International Publishing.

Guo, H., Johnson, M. S., Ercikan, K., Saldivia, L., & Worthington, M. (2024). Large-scale assessments for learning: A human-centred AI approach to contextualizing test performance. *Journal of Learning Analytics, 11*(2), 229–245. https://doi.org/10.18608/jla.2024.8007

Huang, Y., Palermo, C., & Wilson, J. (2026). Accuracy and fairness of generative AI in automated essay scoring: Comparing GPT-4o, feature-based models, and human raters. *Assessing Writing, 69*, 101047. https://doi.org/10.1016/j.asw.2026.101047

Kasneci, E., Sessler, K., Küchenmann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kassir, H. (2023). *The fifth industrial revolution: Unlocking the potential of AI for small and medium sized businesses*. Independently Published.

Korzynski, P., Mazurek, G., Krzypkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review, 11*(3), 25–37. https://www.ceeol.com/search/article-detail?id=1205908

Layton, N. (2024). *What is AI? The curious kid's guide to artificial intelligence*. Hachette Children's Group.

Liew, P. Y., & Tan, I. K. T. (2025). On automated essay grading using large language models. In *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence (CSAI 24)* (pp. 204–211). Association for Computing Machinery. https://doi.org/10.1145/3709026.3709030

Mathew, J. G., Taher, S., Kundu, A., & Barbosa, D. (2026). LLMs do not grade essays like humans. *arXiv*. https://doi.org/10.48550/arXiv.2603.23714

Moorthi, M., Devi, K., Vimala, Manjula, V., & Pattewar, T. (2023). *ChatGPT: Comprehensive study on generative AI tool*. Academic Guru Publishing House.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, 100041. https://doi.org/10.1016/j.caeai.2021.100041

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Lawrence Erlbaum Associates Publishers.

Salinas, A., & Morstatter, F. (2024). The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 4629–4651). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.275

Schaller, N. J., Ding, Y., Horbach, A., Meyer, J., & Jansen, T. (2024). Fairness in automated essay scoring: A comparative analysis of algorithms on German learner essays from secondary education. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 210–221). Association for Computational Linguistics.

Sharma, K., & Bhargav, Y. (2022). Introduction to artificial intelligence and problem solving. *International Journal of Scientific Research & Engineering Trends, 8*(2), 1110–1115.

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Stirling, A. (2024). *The prompt engineer's handbook: Unlocking wealth through words and images*. Mental Health Publishing.

Susilo, A., & Firdaus, M. (2025). A review of artificial intelligence in language assessment: Trends, challenges, and opportunities. *International Journal of Research in Education, 5*(2), 437–446.

University of Technology and Applied Sciences. (n.d.). *Foundation program*. https://www.utas.edu.om/en-us/Study/Foundation-Program

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*, 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Yang, K., Raković, M., Li, Y., Guan, Q., Gašević, D., & Chen, G. (2024). Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence* (pp. 22466–22474). Association for the Advancement of Artificial Intelligence. https://arxiv.org/abs/2401.05655