AI-assisted marking: Functionality and limitations of ChatGPT in written assessment evaluation

Joan Li
Academy for Medical Education, Faculty of Medicine, The University of Queensland

Nikhil Jangamreddy
Faculty of Engineering, Architecture & Information Technology, The University of Queensland

Ruchita Bhansali
School of Public Health, Faculty of Medicine, The University of Queensland

Ryuto Hisamoto
Faculty of Business, Economics & Law, The University of Queensland

Luke Zaphir
Institute for Teaching and Learning Innovation, The University of Queensland

Amalie Dyda
School of Public Health, Faculty of Medicine, The University of Queensland

Mashhuda Glencross
Faculty of Engineering, Architecture & Information Technology, The University of Queensland

Generative artificial intelligence technologies, such as ChatGPT, bring an unprecedented change in education by leveraging the power of natural language processing and machine learning. Employing ChatGPT to assist with marking written assessment presents multiple advantages including scalability, improved consistency, eliminating biases associated with human subjectivity. This work aimed to evaluate the usefulness, reliability and accuracy of ChatGPT in marking written assessments of varied types and to identify its limitations and challenges. ChatGPT was instructed using a set of prompts to mark the assessment based on a rubric. ChatGPT was able to evaluate and assess both coding and reflective assessments and to distinguish between assignments of different quality, demonstrating high consistency and accuracy for higher quality assessments, comparable to a human marking. ChatGPT was also able to generate textual detailed justifications based on the rubric and assessment task description. There was a significant difference in the outcomes generated by different prompts. These preliminary findings suggest that utilising ChatGPT as a marking assistant can increase written assessment marking efficiency, reduce cost and potentially decrease the unfairness and bias by providing a moderating perspective.

Implications for practice or policy:

- Assessment designers could reconsider the design, purpose and objectives of written assessments and leverage ChatGPT effectively for teaching and learning.
- Assessors might consider adapting the technology as a grading aid, to support a human-in-the-loop grading process, providing additional resources and time, moderating and refining individual feedback, to increase consistency and quality.
- Curriculum and programme leaders could develop guidelines around the ethical use of generative AI-assisted assessment practice, monitor and regulate the ongoing evaluation and refinement.

Keywords: higher education, written assessment, marking, feedback practice, generative AI

Introduction

Written assessments test and develop a range of essential skills and competences in higher education including assessing student understanding, promoting critical thinking, developing communication skills and providing feedback for improvement (Price et al., 2011). Traditionally, written assessments are graded by either the academics who design and develop them or by casual academics paid for the specific tasks. The process is both time-consuming and practically challenging to do at scale for a large cohort with limited resources (Wood & Henderson, 2010). It is also cognitively demanding to maintain the standard and quality, ensure consistence and fairness and address individual biases in marking associated with human subjectivity (McConlogue, 2012), even before the pedagogical tasks of providing detailed and implementable feedback that students can use to improve their work.

Generative artificial intelligence (GenAI) technologies, such as ChatGPT (Brown et al., 2020), potentially mark a significant change in education. Its ability to generate text for different purposes, in a wide range of styles, with remarkably greater precision, detail and coherence, offers unprecedented opportunity for both educators and students to enhance teaching and learning though interaction and exploration (Bond et al., 2024; Khosravi et al., 2023). This technology leverages the power of natural language processing and machine learning, enabling users to engage in seemingly natural, open-ended dialogues. These systems then draw upon a vast online knowledge base to predict the most relevant and probable responses (Lund & Wang, 2023). Employing GenAI tools like ChatGPT to assist with written assessment marking could present multiple advantages such as improving consistency, eliminating biases associated with human subjectivity, generating feedback for large cohort, as well as assisting with a faster grading process and enhanced accuracy via repeated evaluation, that is, increase the number of times the same assessment is marked. The use of GenAI may also help to standardise the marking process, which is often impacted not only by subjectivity but also by the experience and knowledge of individual markers. By using this technology, which has been trained on pre-established rubrics and exemplars, unfairness and bias may be decreased by ensuring the consistent and objective application of scoring standards (Baidoo-Anu & Ansah, 2023).

However, there are also associated challenges when it comes to using GenAI as a marking or grading tool. ChatGPT and similar large language models (LLMs) can generate incorrect information (Ji et al., 2023), which raises questions about its accuracy in a given task. Though adept at language processing, these models struggle with the subtle nuances of human expression and context. This can lead to obvious failures to process the task, especially with creative or unconventional responses. This is further exacerbated by data quality issues; LLMs such as OpenAI's GPT, Google's Gemini, the Large Language Model Meta AI and others inherit biases and inequalities through the data they are trained on. The reliance of LLMs on data also raises crucial questions about ownership, privacy and transparency (Wu et al., 2024). Moreover, its hallucinations in citations or sources could complicate the problem regarding the ownership of knowledge and work. There are also ethical concerns associated with replacing human educators with AI, particularly the loss of essential roles teachers play in mentoring and fostering critical thinking and curiosity (Köbis & Mehner, 2021). By carefully addressing ethical considerations, developing robust safeguards against bias and fostering a collaborative approach between humans and machines, OpenAI and other LLMs could become a powerful tool for enhancing education.

The use of AI in education has been explored in recent years (Crawford et al., 2023). A large body of studies has been primarily focused on student use of GenAI and how this may impact the integrity and quality of the assessments and subsequently student learning (Chaudhry et al., 2023; Hwang, 2024; Tlili et al., 2023). Whilst there are many perspective pieces on the use of AI in education and conjectures about the use of GenAI in marking, there is little published robust research investigating this use. A recent paper evaluated the efficacy of ChatGPT in marking short-answer assessments in an undergraduate medical programme, with the suggestion that ChatGPT is a viable assistant to human assessment (Morjaria et al., 2024). Another study investigated the validity and reliability of LLMs in grading English as a foreign language student essays. The findings indicated that LLMs demonstrated a reasonable degree of consistency and potential for grading competency, though further adjustments are needed for a more

refined and thorough understanding of essay criteria (Yavuz et al., 2024). ChatGPT's potential to support English as a foreign language teachers' feedback on students' writing was also reported by Guo et al. (2024). Kasneci et al. (2023) has suggested that teachers can use LLMs to create efficiencies for the grading of student work and other writing assignments. However, to date, few (if any) studies have been conducted to directly compare GenAI marking of various written assignments with the scores given by human markers to evaluate the consistency and accuracy of GenAI marking. Hence, this project aimed to evaluate the usefulness, reliability and accuracy of ChatGPT in marking written assessments of varied contents, formats and qualities and to identify limitations and challenges. By comprehensively evaluating ChatGPT's functionality, we hope to guide the progress and refinement of AI-facilitated assessment tools. This design-based research project was guided by the following two research questions:

(1) How consistently can ChatGPT mark written assessments against a pre-defined criteria sheet?
(2) What are the current capabilities of ChatGPT in terms of marking assessment tasks accurately in line with human equivalents?

Methodology

This was a quasi-experimental research project focusing on quantitative research, numerical data collection and statistical analysis (Gopalan et al., 2020). The project (2024/HE000177) was reviewed by Research Ethics and Integrity and is deemed to be exempt from ethics review under the National Statement on Ethical Conduct in Human Research and relevant University of Queensland policy (PPL 4.20.07).

For our experiments, we used a premium version of ChatGPT which is powered by GPT 4.0, a current state-of-the-art LLM, with a temperature setting of 0.5. We undertook a systematic evaluation of ChatGPT's effectiveness in marking written assessments, and the process is illustrated below (Figure 1).

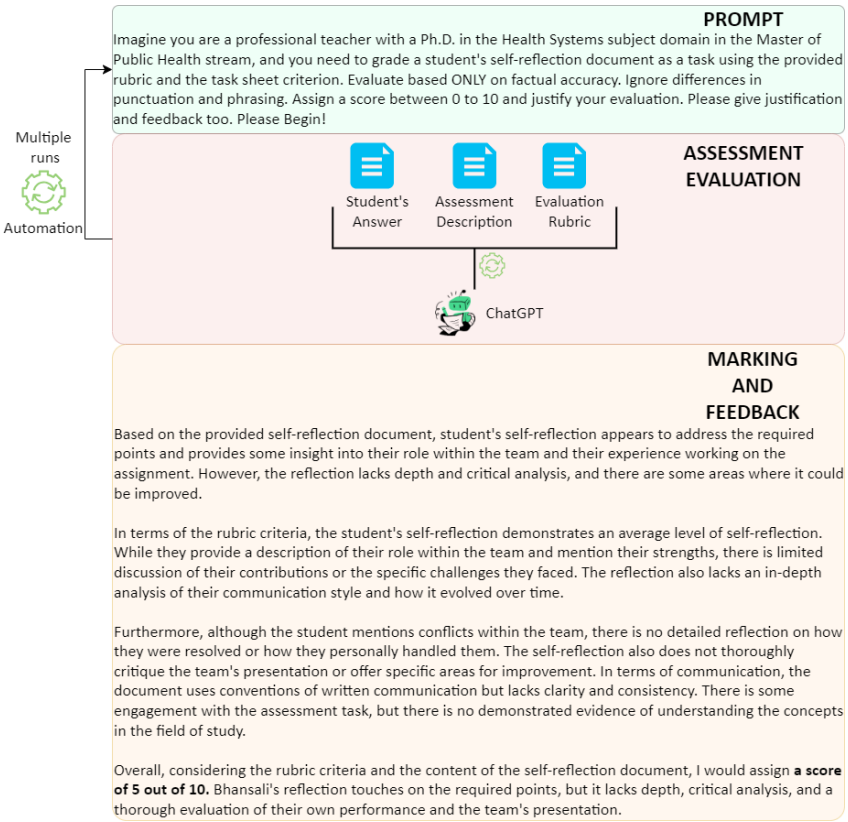

Figure 1. Overview of the methodology

Assessment details

We tested two assessments: a programming assessment from an undergraduate computer science course (first year) and a self-reflective essay from a postgraduate public health course (Health System). In order to investigate the performance of ChatGPT in assessing essays of varying quality, three versions of each assessment representing a poor, an average and a good essay were specifically developed by the students for this study based on task description. The quality of each assessment was checked by a senior member of the research team. Each assessment was also evaluated by human markers with discipline specific knowledge to determine the standard and assigned a score. Marking rubrics were simplified to facilitate ChatGPT processing.

The programming assessment involved the implementation of a 9 x 9 Sudoku puzzle using the Python programming language. We focused only on the readability, structural complexity and documentation of the code since functionality for simple coding assignments is already auto graded using software and specified code assignment rubrics. Two functions, read_board() and print_board(), were chosen as target functions for ChatGPT to analyse for style and documentation. The first function, read_board (), accepts a string as its input and returns a list of lists containing corresponding values in each index. The second function, print_board(), accepts a list of lists as an input and prints the board in a user-friendly manner. For the study, we provided ChatGPT with the style rubric, the coding assignment and the prompt, to assess its ability to refer to the rubric to generate a predictive textual response evaluating the quality of the code.

The public health assessment involved a reflective essay task in which the students were required to write a maximum of 1000-word individual reflection about the content and process of the assessment task, reflect on individual experience of a team assignment within the Health Systems course. The essay provided insights into students' perspectives and thoughts on concepts, emotions, exploration of the topic, team building, the roles undertaken during the assessment and the dynamics of the team. The essay was written in a student-friendly first-person tone. This not only contributed to a valuable learning experience but also helped build students' standpoints in the aspect of assumptions and beliefs. The evaluation procedure included the provision of the reflective assessment, the prompt and the marking rubric, which was modified for ease of use with ChatGPT and focused on overall outcomes. The prompts were developed to be interpretive, instructive and in alignment with the marking details and instructions. Written communication and self-reflection were the two evaluation criteria. The examination sought to examine the students' ability to engage in self-reflection, problem-solving for team-based endeavours and communicating ideas via professional-level written communication.

Rubric details

Compared to the original rubric, the rubric used to test ChatGPT was modified in a way that it retained all the key details and features of the original document and offered increased clarity. With the original rubric, there were significant numbers of violations detected in its initial outputs, possibly due to incorrect processing of the numerical marks according to the marking style. To address the problem, we replaced the fractional marks (eg., 0.2, 0.5) with integer values (0, 1 and 2) to reduce the possibility of ChatGPT misinterpreting the information and awarding marks inappropriately (Table 1). In addition, we made extra modifications in the format, such as simplifying the description for each aspect of the criteria with a single descriptive sentence, changing the file format from PDF to a text file and equalising the mark distributions among all aspects.

Table 1
Coding assignment marking rubric

Aspects	Score	Proficiency level	Description
Programme Structure	0	Good	Layout of the code should ideally use proper vertical whitespacing and horizontal whitespacing.
	1	Satisfactory	
	2	Poor	
Descriptive Identifier Names	0	Good	Code should ideally be written with appropriate identifier names. All non-counter identifier names should follow appropriate naming convention. Code should not use Hungarian notation.
	1	Satisfactory	
	2	Poor	
Named Constants	0	Good	Ideally all non-trivial, fixed values (literal constants) should be represented by informative, well-defined names.
	1	Satisfactory	
	2	Poor	
Single Instance of Logic	0	Good	Blocks of code should ideally not be duplicated in your programme.
	1	Satisfactory	
	2	Poor	
Variable Scope	0	Good	Variables should ideally be declared locally in the function in which they are needed. Global variables should not be used in the code.
	1	Satisfactory	
	2	Poor	
Control Structures	0	Good	Code logic should ideally be structured simply and clearly through good use of control structures using loops and conditional statements.
	1	Satisfactory	
	2	Poor	
In-Line Comment Clarity	0	Good	Comments provide meaningful descriptions of the code. Inline comments ideally should be used appropriately to assist with the readability of the code.
	1	Satisfactory	
	2	Poor	
Informative Docstrings	0	Good	Ideally every function should have a docstring that summarises its purpose. This includes describing parameters and return values (including type information) so that others can understand how to use the function correctly.
	1	Satisfactory	
	2	Poor	

For the reflective essay, initially we considered the rubrics developed by the instructors of the corresponding courses. The original rubric had two criteria and seven numeric scales ranging from 1 to 7 with a percentage assigned to each scale. To reduce the complexity of the rubrics for the LLM and the likelihood of large variability of predictive text generation, we developed simplified rubric based on the original rubrics (Table 2). All subsequent experiments were tested with these modified rubrics.

Table 2
Reflective assessment marking rubric

Learning objectives & criteria	Self-reflection	Communication (written)
7	Comprehensive and critical reflection that provides an in-depth and thorough description of the 3 main reflection points in Part B. Excellent ability to problem solve for future team-based work	Expertly exploits the conventions of written communication at a professional level
6	Detailed critical reflection that provides a thorough description of the 3 main reflection points in Part B. Considerable ability to problem solve for future team-based work	Uses the conventions of the discipline to communicate at a professional level
5	Adequate critical reflection that provides a considered description of the 3 main reflection points in Part B. Good ability to problem solve for future team-based work	Uses the conventions of the discipline to communicate at an effective level
4	Satisfactory reflection that provides a description of the 3 main reflection points in Part B. Sufficient ability to problem solve for future team-based work	Uses some of the conventions of the discipline to communicate appropriately
3	Superficial reflection that provides an insufficient outline of the main reflection points in Part B. Some ability to problem solve for future team-based work	Communicates information or ideas with limited clarity and inconsistent adherence to the conventions of the discipline
2	Limited reflection that inadequately outlines the main reflection points in Part B. Inability to problem solve for future team-based work	Communicates information or ideas in ways that are incomplete, confusing and not appropriate to the conventions of the discipline
1	No demonstration of self-reflection. No ability to problem solve for future team-based work	Some engagement with the assessment task; however, no demonstrated evidence of understanding of the concepts in the field of study

Prompts

Effective prompting can facilitate the desired responses by improving the alignment between user input and AI model output. By crafting well-designed prompts, we can guide LLM systems to generate more accurate, relevant and tailored responses. Initially, a basic prompt instructing ChatGPT to mark the assessment using the rubric was developed, followed by prompt engineering. For the coding assessment, engineered prompts were designed in such a way as to instruct ChatGPT to focus on marking the readability of the code and ignore components related to functionality, which is auto graded. Five different prompts were developed for each assessment and tested by team members with computer science training. All prompts contained a descriptive statement providing relevant context, instructions, and most prompts assigned a role to the AI for the marking task. Each prompt was tested five times initially to analyse the consistency of ChatGPT evaluations. To understand the correlation between marking consistency, accuracy and the number of tests, each prompt was tested a further 15 times (Table 1). Note that each prompt was run as a separate chat, such that each run was independent of other runs. An example of the prompt is shown below:

> You are auto grading a coding assignment. I have provided the following documents: the student's python code, and the assessment rubric to be followed, and you are asked to assign score the student answer based on the evaluation criteria defined in the rubric. Evaluate the student python code based on the assessment rubric. End the assessment with a table containing marks scored in each section along with total marks scored in the assessment. Total marks for the grading is 16. Evaluate based ONLY on factual accuracy. Ignore differences in punctuation and phrasing. Provide the Justification as well.

Evaluation of outcomes

The project adopted a mixed evaluation strategy to assess outcomes across different assessments. The evaluation approach was guided by the quasi-experimental research method focusing on numerical data collection and statistical analysis (Gopalan et al., 2020). We fed the input (assessments and rubric) into ChatGPT, followed by instructive prompts and collected the generated responses for evaluation. Three versions of each assessment (poor quality, average quality and good quality) were assessed by ChatGPT using the modified rubric, under the instruction of specific prompts. We analysed ChatGPT's performance across marking accuracy and consistency, compared to human marking.

The outputs from each prompt were collected and analysed for consistency and accuracy. The quality of the textual justifications was also considered in the evaluation. We evaluated the consistency between each test and consistency across prompts and compared the accuracy against human marker. Five independent outcomes were generated per prompt initially then followed by another 15 runs for each assessment to determine the consistency and accuracy of ChatGPT marking (Table 3). ChatGPT marking reliability data were collected by assessing each essay 100 times using five different prompts (20 times per prompt), at separate intervals using GPT 4.0 model.

Table 3
Test run for each prompt and each assessment

		Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5
Reflective essay	Poor	5 + 15	5 + 15	5 + 15	5 + 15	5 + 15
	Average	5 + 15	5 + 15	5 + 15	5 + 15	5 + 15
	Good	5 + 15	5 + 15	5 + 15	5 + 15	5 + 15
Coding assignment	Poor	5 + 15	5 + 15	5 + 15	5 + 15	5 + 15
	Average	5 + 15	5 + 15	5 + 15	5 + 15	5 + 15
	Good	5 + 15	5 + 15	5 + 15	5 + 15	5 + 15

The final data set for this study comprised marks given by two academic markers from each discipline and parallel scores generated by GenAI using the same rubric. We calculated and compared the descriptive data of 100 independent measurements for each assessment type. Descriptive statistics were carried out to calculate mean, standard error and coefficient of variation of each prompt for each assessment to assess validity and reliability. Two-way analysis of variance (ANOVA) was used to analyse marking consistency between each rating score and each prompt. A comparison between ChatGPT marking and human marking was carried out using a standard t test to assess GenAI marking accuracy. This allowed us to compare and contrast the consistency, reliability and accuracy of GenAI marking compared to human marking, based on the same evaluation criteria.

Results

Evaluation of the consistency of ChatGPT performance in marking written assessments against a pre-defined criteria sheet

When provided with a descriptive rubric, ChatGPT was able to read, review and assess both the reflective essay and the coding assessment. To address the first research question, we conducted a descriptive statistical analysis to evaluate the consistency of ChatGPT marking. We calculated and compared the

descriptive data of 100 independent measurements for each assessment type (Tables 4 and 5). In general, ChatGPT showed higher consistency in marking the reflective writing task than the coding task.

For the reflective task, ChatGPT was able to assign numeric marks accurately to assessments of all qualities based on the rubric. The marks given by ChatGPT were largely consistent with most scores lying between 5.5 and 6.6 for an essay of poor quality, 6.5 to 7.5 for an essay of average quality and 8.4 to 8.9 for an essay of good quality. All five prompts showed acceptable performance in marking with the average coefficient of variation of 13.57% for an essay of poor quality, 10.74% for an essay of average quality and 6.70% for an essay of good quality (Table 4). The decrease in the coefficient of variation suggested that ChatGPT marking consistency increases with the increase in the quality of the assessments. Given the relatively low standard error in all scores, ChatGPT provided a high degree of reliability in marking the reflective essay in alignment with the provided rubric, showing the highest reliability for good quality assessment (SE = 0.13), followed by average quality (SE = 0.17) and poor quality (SE = 0.18).

Table 4
ChatGPT marking of the reflective assessment

Reflective assessment		Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Average
Poor	Mean	6.00	5.50	5.88	6.30	6.60	6.06
	SE	0.16	0.17	0.19	0.12	0.28	0.18
	Coefficient of variation	11.79%	14.15%	14.58%	8.70%	18.65%	13.57%
Average	Mean	6.50	6.66	6.65	6.91	7.55	6.85
	SE	0.16	0.16	0.11	0.18	0.22	0.17
	Coefficient of variation	10.88%	10.64%	7.36%	11.57%	13.23%	10.74%
Good	Mean	8.48	8.40	7.98	8.30	8.93	8.42
	SE	0.11	0.10	0.13	0.12	0.18	0.13
	Coefficient of variation	5.89%	5.24%	6.30%	6.61%	8.85%	6.70%

For the coding assessment, ChatGPT marking was less consistent with significant variations, particularly for the poor and average quality assessments. Across all assessments, the scores given by ChatGPT showed a wider spread of between 4.35 and 8.15 for the assessment of poor quality and 10.75 to 12.75 for the one of average quality. For the assessment of good quality, ChatGPT showed relatively more consistent scoring, 13.65 to 14.65 (Table 5). The average coefficient of variation was 33.89% for the poor quality coding assessment, 15.31% for average quality and 10.53% for the assessment of good quality. A decrease of coefficient of variation from 33.89% to 10.53% showed that ChatGPT performance became more consistent and reliable with the increase in quality of the assessments.

Table 5
ChatGPT marking of the coding assessment

Coding assignment		Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Average
Poor	Mean	4.35	7.55	7.80	6.60	8.15	6.89
	SE	0.33	0.68	0.54	0.57	0.47	0.52
	Coefficient of variation	33.58%	40.42%	31.02%	38.83%	25.58%	33.89%
Average	Mean	11.40	10.75	12.25	11.80	12.75	11.79
	SE	0.41	0.37	0.40	0.34	0.51	0.40
	Coefficient of variation	15.95%	15.35%	14.48%	12.78%	17.97%	15.31%
Good	Mean	13.90	13.65	15.35	14.65	14.30	14.37
	SE	0.38	0.45	0.20	0.23	0.40	0.33
	Coefficient of variation	12.33%	14.90%	5.70%	7.10%	12.65%	10.53%

To summarise, in response to the first research question, these results demonstrated that ChatGPT was able to mark both reflective essay and coding assessment with relative consistency and reliability and capable of distinguishing varying levels of quality and sophistication in the written assessments.

Evaluation of the accuracy of ChatGPT's performance in marking assessment tasks in line with human equivalents

We acknowledge the limitations in recruiting a large number of human markers and the increased possibility of marking variations due to individual differences between the academic markers. Therefore, we applied a marking range instead of the raw scores for comparison. We accept that there will be variability inherent in both the human and the GenAI marking processes, and that GenAI performance will likely differ depending on the prompts and the underlying LLM implementation and how it is trained. To address the second research question and assess the accuracy of ChatGPT marking in line with human equivalents we used a combination of two-way ANOVA and standard t test. By using a two-way ANOVA analysis, we evaluate the consistency of ChatGPT marking while acknowledging the individual differences between different measurements instructed by different prompts.
For the reflective essay, comparison between each measurement across all prompts showed the p value ranging from 0.4 to 0.9 (poor quality to good quality), suggesting that there was no statistical significance overall. However, there were some differences between the performance of different prompts ($p < 0.01$). For the coding assessment, two-way ANOVA analysis showed a significant difference ($p < 0.01$) between the performance of certain prompt, but the overall p value showed no statistical significance (ranging from 0.1 to 0.8, poor quality essay to good quality essay) between each measurement across all prompts. The overall ChatGPT performance for both the reflective and the coding assessments of all qualities showed considerable consistency.

ChatGPT's performance was then compared to human marking. For a poor quality reflective essay, the majority of the ChatGPT marking fell outside of the standard range determined by human markers (Figure 2a). When the test number increased from five to 15 runs, the accuracy and consistency increased for all prompts with 80% of the marking falling within the standard range (Figure 2b). For the coding assessment, when each prompt was tested five times, ChatGPT marking of all assessments showed high degree of inconsistence with significant variation between tests (Figure 2c) and most of the marking did not align with the standard range. An increase in testing number from five runs to 15 runs showed improved performance of all prompts with less than 50% of the marking falling outside of the standard range (Figure 2d). There was also a significant difference in the outcomes generated by different prompts, which could be explained by using ChatGPT's underlying generalised model compared with the use of GitHub co-pilot (Codex LLM), which is better trained for coding. It is likely that LLMs have a much larger training data set of written tasks, allowing them to parse a reflective writing essay more accurately than coding tasks. It is clear from our evaluation that prompts need to be specific and intentional. AI fine-tuning could also be leveraged to pre-train models for specific assessment types appropriate to specific domains.

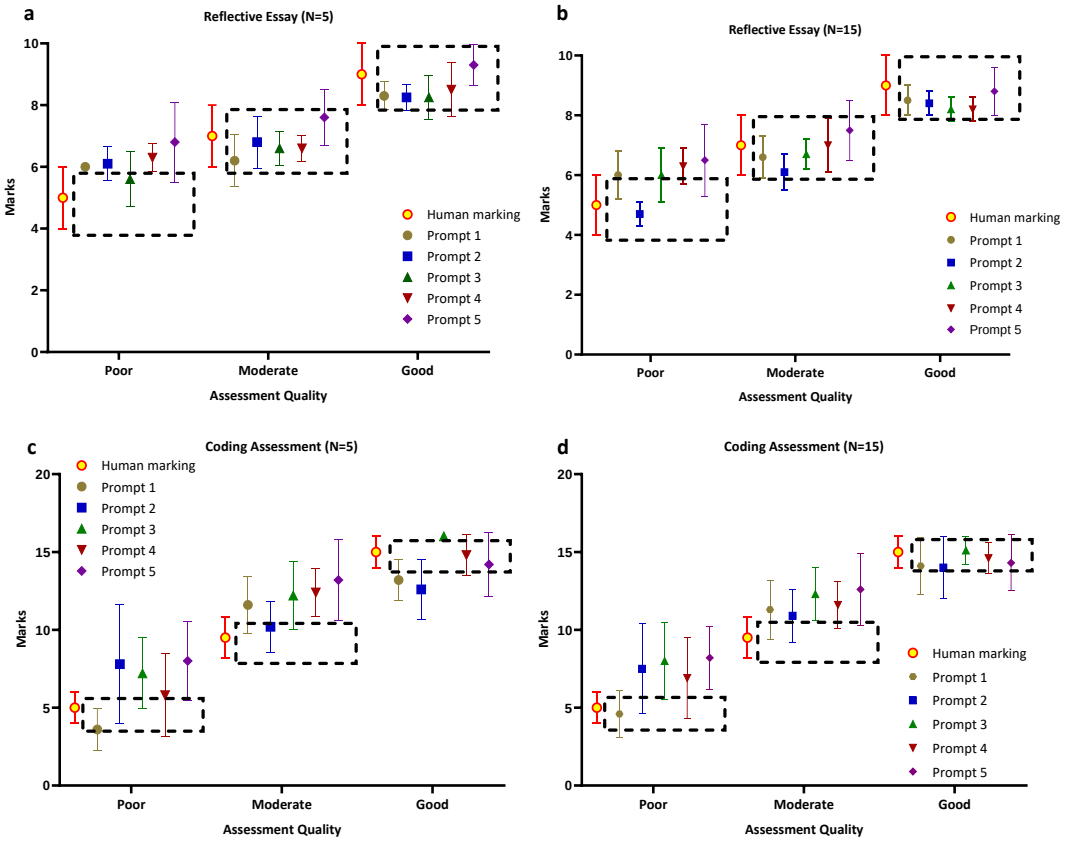

Figure 2. ChatGPT marking of the reflective essay and the coding assessment compared to human marking

A standard *t* test was carried out to directly compare ChatGPT's performance to human marking. Given the large sample numbers, 100 measurements for each assessment, the overall *p* value, $p > 0.99$ for the reflective essay and $p = 0.5$ for the coding assessment, showed no statistical difference between ChatGPT marking and human marking of all assessments (Table 6). ChatGPT marking appeared to be more generous, assigning higher scores to the poor and the average quality assessments. This was especially the case for the coding assessment, but the evaluation of the high-quality assessments was considerably reliable and accuracy (Figure 3).

Table 6
ChatGPT and human overall marking of the coding assessment and the reflective assessment

	Reflective essay			Coding assessment		
	Poor quality	Average quality	Good quality	Poor quality	Average quality	Good quality
Human marking	5–6	7–8	8–10	4–6	8–11	14–16
ChatGPT marking	6.1 ± 0.9 (*n* = 100)	6.8 ± 0.8 (*n* = 100)	8.4 ± 0.7 (*n* = 100)	6.8 ± 2.8 (*n* = 100)	11.8 ± 2.0 (*n* = 100)	14.3 ± 1.7 (*n* = 100)

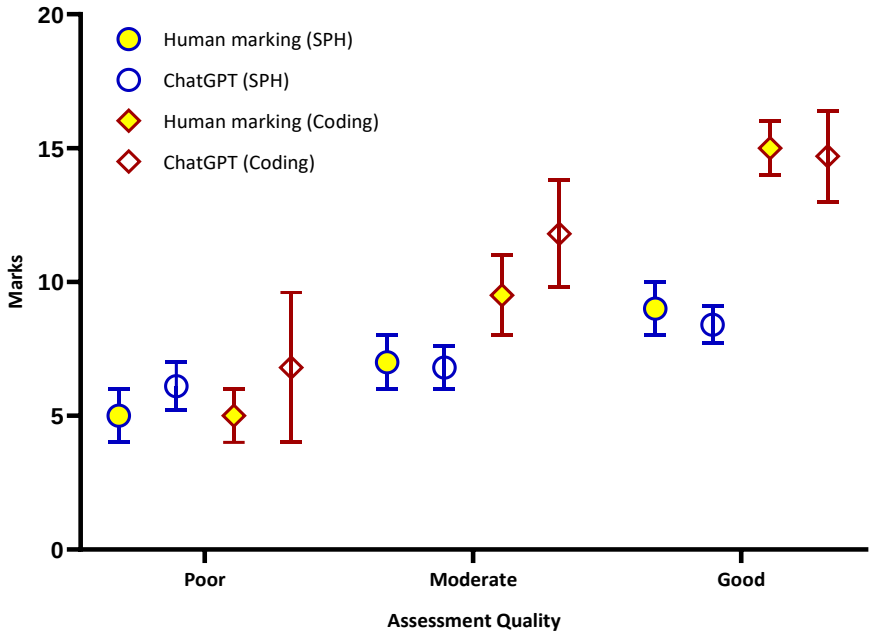

Figure 3. Comparison between human and ChatGPT marking

To summarise, in response to the second research question, ChatGPT demonstrated considerable accuracy in marking both the reflective essay and the coding assessment, comparable to human marking, showing a particular higher degree of confidence in evaluating the reflective essay and in assessments of good quality.

ChatGPT justification of marking

ChatGPT was able to generate detailed textual justifications based on the rubric and the assessment task description when prompted. This required significant prompt engineering through the inclusion of essential criteria in the rubrics as well as revisions of these prompts through testing.

For the reflective essay, justifications focused on two major aspects as suggested in the rubric: self-reflection (evaluation of depth) and written communication (effectiveness and clarity of descriptions). Though there were some shortcomings with the justification provided in terms of reproducibility and transparency, overall, the justifications provided by ChatGPT accurately matched the given scores, indicating that the model was correctly processing the prompts and the resources provided. For example, justification for a good reflective essay with a score of 9 was:

> Demonstrates a good understanding of the team assignment task, effective communication and collaboration skills, and a growth mindset.

In contrast, an average essay scored 6.5 had the following justification:

> However, there is room for improvement in terms of addressing all the learning objectives and enhancing their written communication to align more closely with the conventions of the discipline.

Lastly, for a poorly written essay marked as 5, ChatGPT provided this justification:

> However, it lacks depth and critical analysis of the student's role within the team and their personal growth and learning. The communication in the document is effective but could benefit from a stronger adherence to the conventions of the discipline.

For the coding assignment, justification provided by ChatGPT supporting the scores assigned to individual essay in general, indicated that GenAI could correctly identify problems in the code and process information regarding assessed components (readability, algorithmic logic, documentation). For example, one of the poorly formatted codes received a comment on the inconsistent use of vertical and horizontal spacing, and ChatGPT backed up this justification by adding the following sentence:

> The code uses appropriate spacing and inconsistent at times. For example, there's a lack of spacing around operators and after commas in some places, making the code somewhat less readable.

Moreover, it correctly identified the presence of non-descriptive variable names in the code and awarded 0 marks for the Descriptive Identifier Names section of the rubric with this justification:

> The variable names such as `b`, `r`, `g`, `c` are not descriptive and make the code harder to understand.

Based on the justification provided for each score, ChatGPT showed a capacity in correctly reading and processing the provided information and was able to recognise issues and problems in the assessments according to the rubric. In general, ChatGPT showed higher competency in providing justifications for the reflective essay than the coding assessments. In this project, human markers were not requested to provide justifications for the scores; therefore, there was no comparison.

Effect of prompt on ChatGPT's performance

Our results indicate that ChatGPT does not always consistently process the inputs and the awarded marks do not always align with the justifications provided by the model. This is more frequent for coding assessment than for the reflective essay. There was a difference in the response generated by different prompts which appeared to be independent to the quality of the prompt as the same prompt could generate different responses in the experiment. Regardless of the quality of the prompt, repeated testing could increase the accuracy and consistency of ChatGPT performance, but the prompt difference remained. The quality of the prompts did influence ChatGPT's ability to mark students' assignments and the process, but it is unclear how the quality of the prompt influenced or determined the behaviour of the model and to what degree.

Independent to the prompt the model was given, ChatGPT could review the student's code in a contradicting manner. When asked to assess the same poorly written code, it could award full marks and provide positive comments about the format occasionally; when prompted differently, it would process the code correctly and awarded a partial mark or no mark. In addition, it is evident that sometimes the marks awarded to a certain code did not always match the comments provided by ChatGPT indicating it processed the marking incorrectly. For example, when the model was given an average quality code, although it gave full marks for the algorithmic logic section (6 marks) with positive comments, in the final process of calculating the total mark, it deducted 2 marks from the same section, suggesting the inconsistencies in how ChatGPT processes the given material. In addition, useful comments with examples from a submitted code were not always provided, so relying on the model for insightful comments might expose the system to a risk where some students might receive more useful comments than others despite the similarities in the quality of work they submitted.

Factors influencing ChatGPT's marking accuracy

Though the accuracy and consistency of ChatGPT marking appeared to be independent to the quality of the prompt, other factors such as the type and the quality of the assessments and the number of tests showed positive impact on ChatGPT marking. When we increased the number of times the same files were marked, there was an increase in marking accuracy and consistency, possibly due to the non-deterministic nature of an LLM such as ChatGPT. ChatGPT also showed higher confidence in marking good quality assignments but became less predictable in its performance with poorer quality tasks. Our results

also showed that ChatGPT was better suited to mark the reflective essay than the coding assessment, which are assessments with broad marking categories. It is also possible that the LLMs have a much larger training data set of reflective written tasks, allowing them to evaluate a reflective writing essay more accurately than coding tasks.

Discussion

This study has shown unequivocally that ChatGPT is capable of marking certain types of assessments with a high degree of accuracy, though not necessarily precision. Accuracy in this context is defined as how well the AI's marking applied the marking schema to the students' assessment tasks, as well as matching the overall marks to a human marker. Precision, on the other hand, refers to how wide or narrow the variance is in its marking. This demonstrates that ChatGPT can be used as a marking aid for different types of assessment.

ChatGPT was able to mark reflective writing tasks more accurately and more precisely than the coding assessment. This outcome could be attributed to three factors. First, as an LLM, ChatGPT has been trained on data sets of written tasks (Zuckerman et al., 2023), which may enable it to better analyse reflective writing assessments compared to coding tasks. Second, the reflective writing essay had fewer criteria, potentially reducing the burden on the AI when completing the marking task. GenAI is known to introduce more errors and inaccuracies when tasked with more complex assignments. A third possibility is that the reflective writing task was subjective, unlike the more objective coding task. Marking assessments with objective standards may present challenges for GenAI, which tends to produce inaccuracies despite prompt engineering or instructions (Nikolic et al., 2023). It is likely that all these factors contributed, but without a controlled experiment isolating these variables, it is difficult to determine their relative significance. Simpler marking instructions for tasks with subjective interpretations are likely to produce more reliable and consistent results than tasks that require multiple objective criteria (in ChatGPT).

ChatGPT was also more accurate and precise when marking the "good" and "average" quality work than it does with the "poor" quality assessments. This again could be due to several overlapping reasons. Higher levels of achievement tended to have clearer parameters for those levels of achievement (e.g., a threshold for a high distinction required demonstration of specific skills), making it easier for the model to precisely identify these levels of achievement. Fail-level criteria often have commonalities with low-pass level criteria, making it more difficult for the model to identify. This would seem to be more a function of the overall difficulty of discerning borderline pass/fail assessments when marking (Shulruf et al., 2018), rather than a failure of the model. This could also indicate that the criteria themselves need to be re-evaluated to be clearer for markers.

Recent work investigating the use of ChatGPT to mark short-answer questions shows that there was a difference in the scores provided by ChatGPT with or without the use of a rubric, but these were not significantly different compared to human allocated scores (Morjaria et al., 2024). Rubrics with narrow boundaries or assessments that require answers with low variability such as the coding assessment are likely better assisted with auto grading software. Chat GTP marking is better suited to rubrics with broad marking categories, which is supported by recent work investigating the use ChatGPT to mark written work (Mizumoto & Eguchi, 2023).

Another related possibility is that ChatGPT is being more charitable in its marking than a human counterpart (Mizumoto & Eguchi, 2023). This may make AI marking of fail-level achievement problematic, but also an advantage. Human markers will become fatigued the more they mark, with greater levels of bias against the student due to this cognitive exhaustion. ChatGPT cannot become fatigued, so its marking will remain consistent regardless of the volume of assessment tasks. Regardless, this indicates that ChatGPT cannot be the sole marking apparatus due to its potential inability to discern the poor-quality assessments from the passable. As a marking co-pilot though, these borderline pass/fail cases would be immediately visible to human markers, who would be able to target these marginal assessments for re-evaluation or moderation.

Marking accuracy was improved by increasing the number of times the assessment was marked by ChatGPT and averaging the results. This method may allow ChatGPT to develop a more comprehensive understanding of the material being assessed. By analysing the content from various perspectives and contexts, the system can capture nuances and intricacies that might be missed in a single evaluation. Additionally, averaging the results of multiple runs serves to mitigate any potential errors or anomalies in individual assessments. Or it may be that ChatGPT is learning from variability. Even though ChatGPT is designed for consistency, running the prompts multiple times allows learning from the variability in responses.

There was a significant difference in the outcomes generated by different prompts, which could be explained by the use of ChatGPT's underlying generalised model. Our evaluation also suggests that prompts need to be specific and intentional. The quality and specificity of the prompts can create variance and inconsistency in the marking. This is consistent with previous work which showed that marking accuracy is impacted by the specific prompts input (Webson & Pavlick, 2022). It was also reported that well-crafted prompts yield good responses, while poorly constructed prompts lead to unsatisfactory responses (Heston & Khun, 2023). When ChatGPT is used to assist with marking, the prompts need to be checked against human marking to address this limitation. We designed five prompts and tested each prompt multiple times. Based on our experience, we would recommend testing prompts and using the most accurate and precise prompt for each assessment. Future work should investigate how to develop more sophisticated prompt engineering techniques and understand what makes good prompts to ensure more accurate outputs.

The volatility and the occasional errors within ChatGPT continue to raise concerns about using a language model such as ChatGPT for marking assessments. The project contributes insights into how AI-based technology, specifically ChatGPT in this study, could be used to enhance or improve the marking and feedback processes. This may be a marking co-pilot or other assistance tools for educators. This will be of particular value to courses and programmes with large cohorts, where personalised feedback is impractical and high-quality and consistent marking capacity is nominal. Findings from this study may facilitate innovation in assessment design. By understanding the strengths and limitations of AI-driven marking, educators and technologists could collaborate to enhance the development of AI tools that support more consistent and unbiased marking practices.

These findings need to be considered in the context of the limitations of this study. As this was an initial feasibility assessment, the study was limited to two assessment types, and assessments with descriptive rubric with no numeric scale are yet to be tested. Further exploration of different types of assessment is necessary to gain a better understanding of the full potential of ChatGPT and AI in higher education.

Recommendations

GenAI such as ChatGPT could be considered and utilised as a powerful marking supplement or assistant to human markers, rather than a replacement for the educators. The variability of marking (particularly at the lower end of achievement) means that AI cannot not be relied upon to replace human judgement in terms of marking, but there is a variety of other ways this technology could be used (Bower et al., 2024; Darvishi et al., 2024). One supplement could be as a moderation tool, where the AI's marking is compared against the marking by the human markers to identify any significant discrepancies. Despite GenAI's own biases, its consistency in marking would be useful as a consistent control variable while moderating.

GenAI can also be used to efficiently and effectively provide personalised feedback (Dai et al., 2023; Escalante et al., 2023) based on marking rubrics, which would otherwise be impossible for courses with large enrolments. AI marking is generally accurate (if not necessarily precise), meaning that feedback could be based on the band of achievement rather than the specific mark (e.g., feedback for a student with a high fail, rather than for 47%). In this way, every student would be able to receive personalised

feedback that is timely and relevant. This could also be used as a self-assessment tool, where students could receive feedback before submitting their assessment.

Acknowledgements

We gratefully acknowledge funding provided by the University of Queensland in the form of a student/staff partnership which facilitated this work.

Author contributions

Author 1: Conceptualisation, Investigation, Formal analysis, Writing – original draft, Writing – review and editing; **Author 2:** Investigation, Data curation, Writing – original draft, Writing – review and editing; **Author 3:** Investigation, Data curation, Writing – review and editing; **Author 4:** Investigation, Data curation, Writing – review and editing; **Author 5:** Conceptualisation, Formal analysis, Writing – original draft, Writing – review and editing; **Author 6:** Conceptualisation, Writing – original draft, Writing – review and editing; **Author 7**: Conceptualisation, Writing – review and editing.

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. https://doi.org/10.61969/jai.1337500

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, *21*, Article 4. https://doi.org/10.1186/s41239-023-00436-z

Bower, M., Torrington, J., Lai, J. W. M., Petocz, P., & Alfano, M. (2024). How should we change teaching and assessment in response to increasingly powerful generative artificial intelligence? Outcomes of the ChatGPT teacher survey. *Education and Information Technologies*, *29*, 15403–15439. https://doi.org/10.1007/s10639-023-12405-0

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc. https://doi.org/10.48550/ARXIV.2005.14165

Chaudhry, I., Sarwary, S., El-Refae, G., & Chabchoub, H. (2023). Time to revisit existing student's performance evaluation approach in higher education sector in a new era of ChatGPT — A case study. *Cogent Education*, *10*(1), Article 2210461. https://doi.org/10.1080/2331186X.2023.2210461

Crawford, J., Cowling, M., Allen, K.-A., Crawford, J., Cowling, M., & Allen, K.-A. (2023). Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). *Journal of University Teaching & Learning Practice*, *20*(3). https://doi.org/10.53761/1.20.3.02

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y.-S., Gašević, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. In M. Chang, N.-S. Chen, R. Kuo, G. Rudolph, G. D. Sampson, & A. Tlili (Eds.), *Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies* (pp. 323–325). IEEE. https://doi.org/10.1109/ICALT58122.2023.00100

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education*, *210*, Article 104967. https://doi.org/10.1016/j.compedu.2023.104967

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*, Article 57. https://doi.org/10.1186/s41239-023-00425-2

Gopalan, M., Rosinger, K., & Ahn, J. B. (2020). Use of quasi-experimental research designs in education research: Growth, promise, and challenges. *Review of Research in Education*, *44*(1), 218–243. https://doi.org/10.3102/0091732x20903302

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, *29*(7), 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Heston, T., & Khun, C. (2023). Prompt engineering in medical education. *International Medical Education*, *2*(3), 198–205. https://doi.org/10.3390/ime2030019

Hwang, J. Y. (2024). Artificial intelligence-based self-feedback on medical counselling performance. *Medical Education*, *58*(2), 268–268. https://doi.org/10.1111/medu.15279

Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., & Fung, P. (2023). Towards mitigating LLM hallucination via self reflection. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1827–1843). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.123

Kasneci, E., Sessler, K., Küchemann, 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., … 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

Khosravi, H., Viberg, O., Kovanovic, V., & Ferguson, R. (2023). Generative AI and learning analytics. *Journal of Learning Analytics*, *10*(3), 1–6. https://doi.org/10.18608/jla.2023.8333

Köbis, L., & Mehner, C. (2021). Ethical questions raised by AI-supported mentoring in higher education. *Frontiers in Artificial Intelligence*, *4*, Article 624050. https://doi.org/10.3389/frai.2021.624050

Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Library Hi Tech News*, *40*(3), 26–29. https://doi.org/10.1108/LHTN-01-2023-0009

McConlogue, T. (2012). But is it fair? Developing students' understanding of grading complex written work through peer assessment. *Assessment & Evaluation in Higher Education*, *37*(1), 113–123. https://doi.org/10.1080/02602938.2010.515010

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), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

Morjaria, L., Burns, L., Bracken, K., Levinson, A. J., Ngo, Q. N., Lee, M., & Sibbald, M. (2024). Examining the efficacy of ChatGPT in marking short-answer assessments in an undergraduate medical program. *International Medical Education*, *3*(1), 32–43. https://doi.org/10.3390/ime3010004

Nikolic, S., Daniel, S., Haque, R., Belkina, M., Hassan, G. M., Grundy, S., Lyden, S., Neal, P., & Sandison, C. (2023). ChatGPT versus engineering education assessment: A multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education*, *48*(4), 559–614. https://doi.org/10.1080/03043797.2023.2213169

Price, M., Carroll, J., O'Donovan, B., & Rust, C. (2011). If I was going there I wouldn't start from here: A critical commentary on current assessment practice. *Assessment & Evaluation in Higher Education*, *36*(4), 479–492. https://doi.org/10.1080/02602930903512883

Shulruf, B., Adelstein, B.-A., Damodaran, A., Harris, P., Kennedy, S., O'Sullivan, A., & Taylor, S. (2018). Borderline grades in high stakes clinical examinations: Resolving examiner uncertainty. *BMC Medical Education, 18*, Article 272. https://doi.org/10.1186/s12909-018-1382-0

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, *10*, Article 15. https://doi.org/10.1186/s40561-023-00237-x

Webson, A., & Pavlick, E. (2022). *Do prompt-based models really understand the meaning of their prompts?* In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2300–2344). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.167

Wood, E. H., & Henderson, S. (2010). Large cohort assessment: Depth, interaction and manageable marking. *Marketing Intelligence & Planning*, *28*(7), 898–907. https://doi.org/10.1108/02634501011086481

Wu, X., Duan, R., & Ni, J. (2024). Unveiling security, privacy, and ethical concerns of ChatGPT. *Journal of Information and Intelligence*, *2*(2), 102–115. https://doi.org/10.1016/j.jiixd.2023.10.007

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2024). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13494

Zuckerman, M., Flood, R., Tan, R. J. B., Kelp, N., Ecker, D. J., Menke, J., & Lockspeiser, T. (2023). ChatGPT for assessment writing. *Medical Teacher*, *45*(11), 1224–1227. https://doi.org/10.1080/0142159X.2023.2249239

Corresponding author: Joan Li, j.li1@uq.edu.au

Please cite as: Li, J., Jangamreddy, N., Bhansali, R., Hisamoto, R., Zaphir, L., Dyda, A., & Glencross, M. (2024). AI-assisted marking: Functionality and limitations of ChatGPT in written assessment evaluation. *Australasian Journal of Educational Technology*, *40*(4), 56–72. https://doi.org/10.14742/ajet.9463