Evaluating AI-Generated Feedback for Formative Assessment in Higher Education

Xiaolei Li, and David Chen (https://orcid.org/0000-0001-8690-7196)
Griffith University, Queensland, Australia

Abstract

The consistent provision of effective formative feedback on assessment is an integral part of the learning process, yet it is one of the most demanding and challenging aspects of learning and teaching, especially when required at scale. It is therefore unsurprising that amid the exploration of AI tools in teaching and learning, there is growing interest in the application of Large Language Models (LLMs) to support formative feedback in the learning process. Using student submissions from a course that included structured peer review activities, the study reported in this paper is a comparative analysis of AI-generated feedback and scoring (GPT-4). The findings show AI-generated feedback provides structured and detailed comments for assessment components based on clear and objective criteria, with strong agreement with human marking practices. Limitations appeared in areas requiring subjective judgment, such as evaluating the quality of peer reviews and the depth of student reflection. In these cases, human educators provide more nuanced interpretations based on contextual and pedagogical understanding. Findings put emphasis on the importance of human oversight for qualitative and interpretive evaluation. These findings suggest that a balanced human-AI approach, grounded in pedagogical intent and careful integration, is essential for the effective use of AI-assisted feedback in higher education.

Keywords: AI-generated feedback; AI-assisted Assessment; Formative Assessment; Peer Review; Higher Education

1. Introduction

Formative assessment is crucial in education, acting as a continuous feedback loop to support student learning and instructional practices. It provides information to students and educators

about current performance relative to learning goals, enhancing learning experiences and teaching practices (Black & Wiliam, 2009; Irons & Elkington, 2021;). Black and Wiliam (2009) demonstrated the transformative effects of formative assessment, but these effects are only possible when ongoing feedback significantly enhances teaching practices and student performance. Irons and Elkington (2021) further emphasize the importance of integrating formative feedback into daily teaching practice to support student engagement and provide actionable guidance for both students and educators.

Timely and personalized feedback is difficult to sustain in large classes, where workload pressure often limits the depth and consistency of educator input (Boud & Molloy, 2013; W. Dai et al., 2023). Empirical studies suggest that automated assessment systems can improve consistency, efficiency and scalability by supporting routine feedback processes (e.g., Beerepoot (2023); Y.-C. Lee & Fu, 2019))

Given the dialogic nature of feedback, it is the exploration of large language models (LLMs) that are attracting the most attention. The authors, motivated to contribute to the exploration of AI for facilitating effective feedback at scale, and with a particular interest in peer feedback, set out to investigate the effectiveness of ChatGPT 4.0 (one typical LLM) as a tool for effective and efficient formative feedback on assessment in group learning activities where peer feedback is part of the learning process. Specifically, the authors were interested in how AI-generated feedback compares with human feedback across assessment criteria in an authentic university setting, gaining insights into acceptance and into how LLMs might integrate with human to facilitate feedback that is effective and at scale.

The study (Ethics approval GU Ref No: 2023/755). adopts an exploratory approach to examine LLM-generated feedback in a structured peer-review context within a web application development course. The broader study context is the application of LLMs to

formative assessment feedback in higher education, particularly in low-stakes assessment settings where scalability and consistency are ongoing challenges. To support analysis of feedback structure, the study draws on the feedback model proposed by Hattie and Timperley (2007) as an analytic lens for examining how feedback is oriented and framed. This model is used descriptively to characterize the focus of AI-generated feedback, rather than to evaluate feedback effectiveness or learning outcomes.

GPT-4 was selected as a representative instance of contemporary LLMs, based on its availability and widespread use at the time of the study, as well as its demonstrated capacity to generate structured, rubric aligned textual feedback. The study does not seek to evaluate GPT-4 as a product, but rather to use it as an illustrative example for examining the behavior of LLM-based feedback systems more generally. Aware of the importance of prompt design in shaping LLM outputs, we applied the C.R.E.A.T.E. (Birss, 2023) and C.L.E.A.R. (Lo, 2023) prompt design frameworks to construct prompts aligned with assessment rubrics. This approach is consistent with recent work emphasizing that the pedagogical value of LLM-generated feedback depends less on the model itself and more on how educators design, constrain, and iteratively evaluate AI outputs using structured prompts and professional judgment (Correia et al., 2025). These frameworks together demonstrate how general prompt design principles can be applied to the settings of formative assessment feedback.

The study employed a mixed-methods approach using comparative content analysis of anonymized student submissions from a course featuring structured peer-review exercise. AI-generated scores and comments were compared with those produced by human markers under shared rubric conditions. GPT-4 was applied post hoc to submissions from a completed course and was not integrated into live teaching or used to provide direct feedback to students. As a result, the study does not examine student uptake of feedback or learning

outcomes. Given the small sample drawn from a single course context, the findings are intended to be exploratory rather than generalizable, with a focus on identifying patterns of alignment and divergence between AI-generated and human feedback.

The study is guided by the following research questions:

(1) How can GPT-4 be configured and applied under controlled conditions, to generate rubric-aligned formative feedback on peer review submissions?

(2) How does GPT-4-generated feedback and scoring compare with human marker feedback scoring across different assessment components?

Rather than validating model performance or claiming pedagogical effectiveness, the study documents observed feedback behaviors that may inform future research on AI-human collaboration, prompt design, and the integration of AI-assisted feedback into formative assessment practices.

The study makes theoretical, practical and future research contributions in that it: Examines AI-generated as a design and alignment problem by analysing how feedback characteristics differ across objective and subjective assessment criteria under shared rubric constraints; Provides a comparative analysis of AI-generated and human feedback on the same peer-review submissions, identifying patterns of alignment and divergence that may inform rubric design, prompt construction, and teacher calibration;

Offers preliminary evidence to support future research on live integration and AI-human collaboration in formative feedback practices.

2. Literature Review

2.1 Principles of Effective Feedback

Nicol and Macfarlane-Dick (2004) presented seven principles of effective feedback, emphasizing clarity, constructiveness, and support for student self-assessment. These principles frame feedback as an iterative process that helps students understand their current performance and plan for future improvements.

Hattie and Timperley (2007) further conceptualized feedback through three guiding questions, “Where am I going?”, “How am I going?”, and “Where to next?”, and four levels of feedback focus, task, process, self-regulation, and self. This model highlights the structured and multidimensional nature of feedback and has been widely used to analyze how feedback is oriented and framed in educational settings. In the present study, these three guiding questions are used as an analytic lens to characterize the structure and orientation of AI-generated feedback, rather than as a framework for evaluating feedback effectiveness or learning outcomes.

2.3 Theoretical Perspective Relevant to Feedback

From a sociocultural perspective, learning is mediated through interaction with others and with tools (Vygotsky, 1978). Feedback, particularly in peer-review settings, can be understood as a form of social mediation that supports learners’ engagement with standards, expectations, and disciplinary practices. This perspective is relevant when considering AI-support feedback, as AI tools may function as mediating artifacts within learning activities rather than as independent instructional agents.

Building on this perspective, (Henderson et al., 2019) highlighted persistent challenges in feedback practice in higher education, particularly the disconnect between feedback provision and student uptake. They argued that feedback often fails to achieve its intended impact not because of insufficient quality or detail, but because students struggle to interpret, trust, and act on feedback. From this view, effective feedback depends on learning designs that actively engage students in sense-making, reflection, and decision-making processes, rather than positioning them as passive recipients of comments. This emphasis on student agency and feedback use is especially relevant in peer review settings and informs the present study’s focus on feedback structure and alignment rather than feedback outcomes.

Zimmerman's (2008) Self-Regulated Learning (SRL) framework further emphasizes the role of feedback in supporting learners' abilities to monitor, evaluate and adjust their learning strategies. Feedback that clarifies goals and provides information about current performance can support self- regulation; however, the effectiveness of such support depends on how feedback is interpreted and used by learners. Although our study focuses on educator-facing use of AI for assessment and feedback generation, this framework provides a conceptual background for understanding why the structure and orientation of feedback are central to formative assessment. Prior research also suggests that AI tools may support aspects of SRL in online learning environments (e.g., Wong et al. 2019).

Zimmerman's (2008) Self-Regulated Learning (SRL) framework further emphasizes the importance of learners' active role in monitoring, evaluating, and regulating their own learning. Feedback is a key mechanism that supports SRL by helping learners set goals, assess progress, and adjust strategies. In the context of online learning and performance-based tasks, timely, specific feedback is particularly important because it enables students to identify gaps between current performance and desired outcomes and to make improvements through iterative practice.

2.4 AI in Feedback on Assessment Mechanisms

Recent studies have increasingly focused on the application of AI to feedback on assessment mechanisms. Research suggests that AI systems can support the generation of rubric-aligned feedback, enhance consistency across markers, and reduce assessment related workload in large classes (Abdel Aziz et al., 2024; Ballantine et al., 2024). Chang et al. (2023) examined both opportunities and challenges associated with AI adoption in education, emphasizing the importance of instructional alignment and learner support.

AI performance varies substantially depending on task structure and the nature of assessment criteria. For example, Ali et al. (2023) found that ChatGPT performed well in structured

assessment tasks but struggled with tasks requiring critical appraisal and nuanced qualitative judgment. AI-generated feedback is more dependable when assessment criteria are explicit and rule-based, and less reliable when interpretive judgment is required. Such distinctions are particularly relevant in peer-review settings, where assessment often spans both objective and subjective dimensions.

2.5 Prompt Engineering and Feedback Quality

Research on LLM behavior demonstrates that model responses are sensitive to how tasks and constraints are specified in prompts, with different prompt formulations leading to differences in output quality and focus (Brown et al., 2020; Reynolds & McDonell, 2021). Recent conceptual work further emphasizes that the pedagogical value of LLM-generated outputs in educational settings depends on structured prompt design, iterative refinement, and sustained teacher oversight, positioning AI-generated feedback as a component of assessment design rather than an autonomous instructional agent (Correia et al., 2025).

In higher education research, recent review studies have identified prompt engineering as an important consideration for aligning generative AI outputs with educational goals, while also reporting wide variation in how prompts are designed and applied across studies (D. Lee & Palmer, 2025). In the context of formative feedback, (W. Dai et al., 2023) demonstrated that t AI-generated feedback guided by rubric-based prompts can produce readable, criteria-referenced comments, but also revealed variability in alignment with instructor judgments across assessment dimensions.

3. Methodology

3. 1 Research Design

In addressing the challenges of delivering timely and detailed feedback identified in our literature review, our study employed a comparative content analysis design to examine similarities and differences between AI-generated feedback and human feedback under shared assessment conditions. The analysis focused on feedback produced for structured peer

review tasks, treating AI-generated feedback as an artifact of prompt design and rubric alignment rather than as a deployed instructional intervention.

GPT-4[1] was selected as a representative example of modern large language models available at the time of the study. Its use enabled an examination of how a state-of-the-art LLM could be configured, through prompt design and rubric specification, to generate formative feedback comparable to that of human markers. Importantly, the study does not seek to validate GPT-4 as an assessment tool, but rather to explore patterns of alignment and divergence between AI-generated and human feedback under controlled conditions.

To ensure comparability, both human evaluators and GPT-4 were provided with the same detailed assessment rubric outlining criteria for evaluating student submissions. For human evaluators, regular inter-evaluator reliability checks were conducted to ensure scoring consistency, and any differences were addressed through brief calibration discussions. For GPT-4, student submissions were uploaded one at a time via the API, without model retraining, and no changes were observed in the AI's analysis over time, suggesting minimal impact from potential learning effects. This standardized approach, including fixed prompt instruction (as shown in Figure 1), API-based input, and the same rubric for all evaluations, maintained uniform conditions, minimized subjectivity, and enabled both human and AI feedback to be assessed under consistent and comparable criteria.

We adopted an exploratory, pre-integration comparative approach, focusing on how GPT-4 can be configured to generate rubric-aligned formative feedback and how such feedback compares descriptively with established human marking practices.

3.2 Data Collection

[1] OpenAI. (2023). GPT-4-0613. Retrieved from https://www.openai.com/research/gpt-4

The study analyzed de-identified peer review submissions from a Web Application Development course at Griffith University during Trimester 2, 2023. This course enrolled a cohort of 131 undergraduate and postgraduate ICT students. This course is a comprehensive web programming course that covers server-side programming with PHP, database interaction (CRUD), templating, MVC framework, and security, providing students with practical skills to build robust and secure dynamic web applications.

Students completed seven peer reviews for formative assessment across the trimester, each contributing 2% to the course grade; however, only the best five scores were counted, resulting in a total weighting of 10%. This low-stakes design limited the impact of individual feedback variations on students' overall performance.

3.3 Peer Review Scenario

The course implemented a structured peer review process consisting of three stages:

Group Formation: Initially, students were randomly grouped by the teacher in the workshop, with at least 3 students per group, to foster a diverse and comprehensive peer review process.

Peer Review Activity: Students participated in a peer review mechanism. Each student took turns demonstrating their code/solution for that week's exercise to the group, and then their peers in the group provided written feedback to that student on the demonstration, guided by a peer review template provided by the teacher Each student was required to review at least two other students and received feedback from at least two peers.

Feedback Evaluation and Reflection: Each student then wrote a reflection on feedback from their peers in their group, concentrating on identifying strengths and areas for improvement.

To clarify the marking scheme: 0.5 marks are awarded for attempting the exercise and participating in the demonstration, 1 mark for writing the review, and 0.5 marks for writing the reflection.

3.4 Use of GPT-4 and Prompt Design

To guide the generation of AI-based feedback, prompt design followed prior work that employed rubric-aligned prompts to for assessment tasks (W. Dai et al., 2023). Two complementary prompt-engineering frameworks were used to support clarity, consistency, and alignment between AI-generated feedback and course assessment criteria.

First, we drew on Lo's (2023) C.L.E.A.R. framework for prompt engineering, which highlights the importance of creating concise, logical, explicit, adaptive, and reflective prompts. Using this framework, we developed and iteratively refined prompts reflecting the assignment rubrics, focusing on “Peer Review”, “Submit and Demonstrate Code”, and “Reflection”.

To further strengthen the structure and clarity of our prompts, we applied the C.R.E.A.T.E. framework (Birss, 2023), which outlines six key elements: Character, Request, Examples, Additions, Type of Output, and Extras (Figure 1). This framework provided practical guidance for constructing detailed, content-aware prompts that communicate the intended assessment standards to the AI.

While GPT-4 cannot replicate contextual judgment of human educators, the combined use of these frameworks was intended to increase the alignment between AI-generated feedback, course learning objectives, and established marking practices.

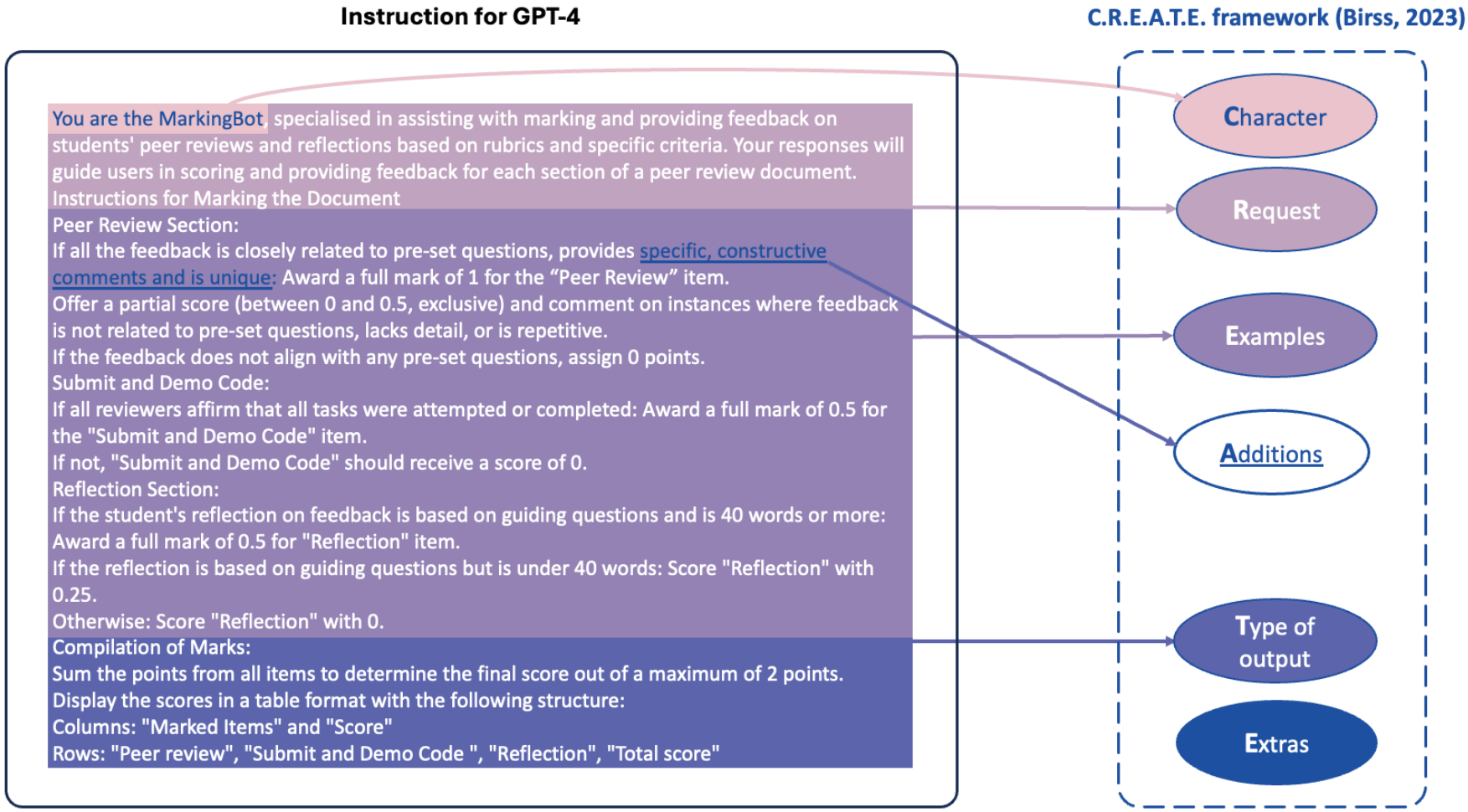

Figure 1 Instruction for GPT-4 Using C.R.E.A.T.E. Framework (Birss, 2023)

Detailed instructions for GPT-4, using the C.R.E.A.T.E. framework (Birss, 2023), are shown in Figure 1.

Character: GPT-4 was instructed to act as an automated marker applying predefined assessment criteria in evaluating student submissions.

Request: GPT-4 was asked to assess peer reviews and reflections written by students, assign scores and provide detailed feedback based on predefined criteria for each section of the student submissions.

Examples: Detailed scoring examples were provided, such as full marks for well-aligned and constructive feedback, with reduced or no points for missing criteria.

Additions and Type of Output: The instruction could benefit from additional guidance on edge cases, while the output is clearly defined as a structured score table to enhance transparency and understanding.

Extras: Although the "Extras" category is part of Birss's (2023) C.R.E.A.T.E. prompt engineering framework, it is distinct from the C.R.E.A.T.E. employability skills framework. This factor was not included in our prompt this time but may be considered in future studies to further develop our AI-assisted assessment tool.

3.5 Sampling and Selection for GPT-4 Trial

For the GPT-4 trial, a stratified sampling approach was used to select a representative sample of student submissions that reflected common performance patterns observed in the course. This sample included submissions that received full marks as well as those with specific deductions, encompassing both actual and simulated student submissions, to ensure a comprehensive evaluation of GPT-4's performance across a variety of potential assessment challenges.

Our initial analysis showed that approximately 68% of the student submissions achieved full marks. While this high percentage reflects strong overall student performance, it presents a

potential ceiling effect from a research perspective, limiting variation for comparative evaluation. To address this, we identified common performance categories based on patterns observed during the initial review of the dataset. In addition to the “Full Mark” cases, we noted three recurring types of underperformances: (1) Failure to Attempt Tasks, where students missed required components; (2) Poor Quality of Peer Review Feedback, where responses lacked detail or failed to address the guiding questions; and (3) Inadequacies in Self-Reflection, where reflections were superficial or misaligned with task expectations.

To ensure sufficient performance variance, we included a balanced sample of submissions with common deductions. Specifically, we randomly selected 10 cases from the “Full Mark” category, representing the most frequent outcome, to allow for analysis of GPT-4’s handling of high-quality work. From the less frequent categories “Failure to Attempt Tasks,” “Poor Quality of Peer Review Feedback,” and “Inadequacies in Self Reflection,” we selected 5 cases each, totaling 15 deduction cases. This sampling strategy ensured coverage of common performance issues, as outlined in Figure 2.

Cases of late submissions and incorrect formatting were excluded from our trial.

We included two simulated scenarios to test GPT-4's handling of potential edge cases, such as suspected plagiarism or inappropriate content. Although such cases were not observed in this course offering, they were included to explore how AI-generated feedback might respond to situations that could arise in future implementations.

This sampling approach supports an exploratory comparison of AI-generated and human feedback across a range of commonly observed submission conditions, rather than a comprehensive evaluation of overall performance or generalizability.

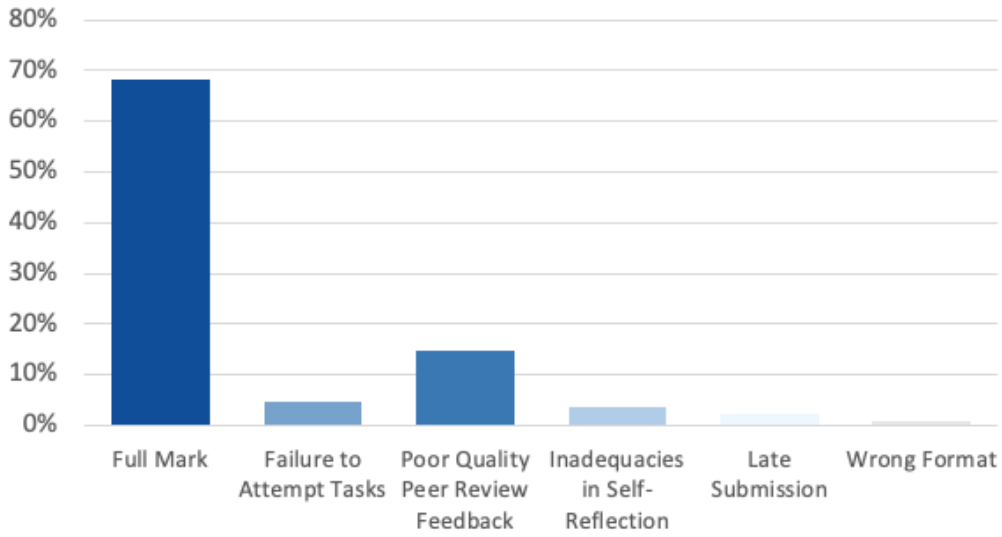

Figure 2 Distribution of Categories in Student Submissions for the Web Application Development Course

3.6 Data analysis

Building on the selection and categorization processes described, the data analysis focused on comparing the feedback and scores provided by GPT-4 with those given by human educators for a set of categorized peer review submissions from a completed course. These categories, "Full Mark" “Failure to Attempt Tasks,” “Poor Peer Review,” and “Inadequacies Self-Reflection”, were used to structure the analysis and examine the extent to which GPT-4’s feedback and scoring aligned with established human marking practices across different types of submission. While student performance after receiving GPT-4’s feedback was not measured in this study, our results provide descriptive insights into how AI-generated feedback compares with human feedback under different submissions conditions.

4. Results

4.1 Full Mark and Non-Attempt Cases

In cases where students achieved full marks, GPT-4’s scores were 100% agreement with that of the human teachers. As shown in Table 1, GPT-4 also generated detailed, criterion-referenced feedback describing how the submission met each assessment requirement, while

the human marker did not provide written comments. This pattern was consistent across full-mark cases.

A similar pattern was observed **in all cases where students failed to attempt tasks. In these cases, both GPT-4 and human markers assigned identical scores, again achieving 100% agreement.** As shown in Table 2, GPT-4 generated detailed feedback explaining the basis for deductions across assessment components. This included identifying missing peer reviews, unaddressed submission requirements, and completed reflection components. These full-mark and non-attempt cases (Table 1 and Table 2) indicate strong alignment between GPT-4 and human markers when assessment outcomes were clearly defined.

Table 1 Comparative Analysis of Sample Feedback for Full Mark Cases by GPT-4 and Human Teachers

Marker	**Score**	**Comments**
Teacher	2	Nil
GPT-4	2	Peer Review Section: The reviewer addressed each pre-set question while providing specific and constructive comments for each reviewee. There were no repetitive comments. Therefore, the reviewer receives the full score for this section. Score: 1 Submit and Demo Code Section: In all the feedback received, all reviewers stated that the tasks were attempted or completed, so the full score is awarded for this section. Score: 0.5 Reflection Section: The student's reflection is based on the given guidelines and is more than 40 words long. The reflection receives the full score. Score: 0.5 Below is the summary of the scores obtained in each section: | Marked Items | Score | |---|---| | Peer review | 1.0 | | Submit and Demo Code | 0.5 | | Reflection | 0.5 | | Total score | 2.0 | The student gets a final score of 2.0 out of a possible 2.0. Well done!

Table 2 Comparative Analysis of Sample Feedback for Failure to Attempt Tasks by GPT-4 and Human

Marker	Score	Comments
Teacher	0.5	Nil
GPT-4	0.5	Peer Review Section: The student did not provide any peer review. Score: 0 Comment: You are encouraged to participate in the peer review process to enhance your learning experience Submit and Demo Code Section: The student did not mention any actions regarding the "Submit and Demo Code" part so this part can't be evaluated. Score: 0 Reflection Section: The student did indeed reflect on the feedback and questions for this part have been answered and the word count is more than 40 words. Score: 0.5 Comment: Your reflection is insightful. Please remember to share your reflections with your teammates; this helps you all learn and grow as a team. Continue to engage in this practice in future activities, it provides a good evaluation of your work. Here's the final table of the marked items: | Marked Items | Score | |---|---| | Peer review | 0 | | Submit and Demo Code | 0 | | Reflection | 0.5 | | Total score | 0.5 | The student needs to participate fully in the workshop activities for better scores in the future.

4.2 Peer Review Assessment

The main area of disagreement between GPT-4 and human teachers was observed in the Peer Review and Reflection section. In assessing the quality of peer review, GPT-4 successfully identified common issues such as missing comments. For instance, as shown in Table 3, GPT-4 noted, “no comments for section 3 and 4” (Table 3: Case 4). However, in four other selected cases where human markers deducted points for being “too brief”, GPT-4 failed to consistently recognize this deficiency shown in Table 3 under Case 1,2, and 3. It was only in one instance that GPT-4 identified that “the feedback lacked specificity and could have been more unique” (Table 3: Case 5). These results suggest that while GPT-4 can identify explicit deficiencies, it sometimes lacks the subjective judgment necessary to fully align with human evaluators. Even when it generally agrees with the human teacher, discrepancies in the application of evaluative criteria can still occur.

Table 3 Comparative Analysis of Feedback Deduction Marks in Peer Review Section: Sample Cases Evaluated by GPT-4 and Human Teachers

Case No.	Marker	Deduction Mark	Comments
1	Teacher	-0.5	Reviews are too brief, need to be more specific and provide more details.
	GPT-4	None	The reviews are closely related to the pre-set questions, give specific constructive feedback, and address each item.
2	Teacher	-0.5	Most of the reviews are too brief and need to be more specific e.g. item 2 needs to describe not just state.
	GPT-4	None	The Peer Review section shows that the student has given constructive feedback closely related to the pre-set questions without being repetitive.
3	Teacher	-0.5	Some of the reviews are too brief and need to be more specific e.g. item 2 needs to describe not just state
	GPT-4	None	The assessments are in full compliance with the pre-set questions. They are insightful, constructive, and unique for each individual reviewed.
4	Teacher	-0.5	Cannot simply write 'No Comment' as that is not really a review.
	GPT-4	-0.25	While there was specific feedback provided for peer S063, peer S066 only received an overall positive response with no comments for section 3 and 4.
5	Teacher	-0.5	Reviews are far too brief and need to be more specific with more details.
	GPT-4	-0.25	The feedback provided was related to the pre-set questions and was quite detailed overall for all reviewees. However, there were some instances where the feedback lacked specificity and could have been more unique (Review for S044 and S086). The response provided for S039 was more specific and complete.

In Table 3, which focused on peer review quality, the differences between human and GPT-4 assessments primarily emerged in cases requiring subjective judgments about feedback constructiveness. For instance, in Table 3: Case 1, human markers deducted 0.5 marks for reviews being "too brief" with comments like "need to be more specific," while GPT-4 awarded full marks, noting the feedback was "closely related to pre-set questions" and "constructive". This pattern was repeated in Table 3: Cases 2 and 3, where human evaluators identified issues with insufficient detail that GPT-4 failed to recognize, demonstrating its current limitations in evaluating the qualitative depth required for effective peer feedback. However, in Table 3: Case 4, GPT-4 partially aligned with human judgment by deducting

0.25 marks for missing comments in specific sections, showing a capacity for detailed analysis when deficiencies were explicit rather than interpretive.

4.3 Reflection Assessment

Reflection deductions provided insights into GPT-4's strict adherence to the course rubric, particularly in managing word counts, as seen in Table 4, where GPT-4 noted "less than 40 words" (Table 4: Case 1). This level of precision, focusing on structural aspects such as word count, was not always matched by human teachers who sometimes prioritized content depth and relevance. For instance, in Case 2, 3, 4 and 5 (Table 4), where teachers identified reflections as vague or insufficiently addressing the required questions, GPT-4 primarily applied deductions based on measurable criteria, such as word count, "it is not 40 words long" (Table 4: Case 2).

In simulated scenarios addressing various reflection issues, GPT-4 was able to identify "unrelated information" (Table 4: Case 6) and "The language used is inappropriate" (Table 4: Case 7), making corresponding deductions. This suggests that GPT-4 possesses the capability to effectively evaluate and respond to language use and content relevance, both of which are critical to the integrity and quality of reflective assessment in education.

Table 4's reflection assessments revealed more nuanced differences, arising from the extended writing nature of this component. While both assessors emphasized rubric compliance, human markers prioritized content substance over structural elements, deducting marks for vague responses such as "Yes" as an answer (Table 4: Case 4) or insufficient future improvement plans (Table 4: Case 5), whereas GPT-4 focused on measurable criteria like word count (Table4: Case 2).

In simulated edge cases, GPT-4 demonstrated unexpected sophistication, identifying inappropriate language use ("frustration rather than reflection", Table 4: Case 7) and

unrelated content (Table 4: Case 6), suggesting stronger performance in detecting overt violations versus subtle qualitative deficiencies. This contrast highlights the complementary strengths of human contextual understanding and AI's capacity for consistent structural analysis. The reflection analysis shows differing emphases between human and AI evaluators: human markers applied more interpretive judgment to content substance, whereas GPT-4 applied rubric-specified structural criteria more consistently.

Table 4 Comparative Analysis of Feedback Deduction Marks in Reflection Section: Sample Cases Evaluated by GPT-4 and Human Teachers

Case No.	Marker	Deduction Mark	Comments
1	Teacher	None	None
	GPT-4	-0.25	The student's reflection is based on the guiding questions and is less than 40 words.
2	Teacher	-0.25	Little brief, should also discuss approaches to improve in the future
	GPT-4	-0.25	The student has provided a reflection based on the feedback. However, it is not 40 words long.
3	Teacher	-0.25	Some of the reflection is vague and unclear. Please try to be more specific and also specifically address the required questions.
	GPT-4	None	The reflection was detailed and based on guiding questions. At over 40 words, it adequately complied with the required word count.
4	Teacher	-0.25	Yes' is not a suitable answer for 'What have you learnt from completing this week's exercise.'
	GPT-4	None	The reflection is based on guiding questions and exceeds the 40 words requirement, indicating thoughtful introspection.
5	Teacher	-0.25	It is important to consider how you are going to improve in the future based on the feedback received. This should be clear and specific.
	GPT-4	None	The reflection is based on the guiding questions, and the word count is significantly more than 40 words.
6	Teacher	-0.5	Unrelated content
	GPT-4	-0.5	Reflection: In the reflection section, the reviewee seems to have shared some unrelated information instead of reflecting on the reviews received. It seems they didn't respond to the guiding questions or fulfil the 40-word minimum.
7	Teacher	-0.5	Use negative/impolite words
	GPT-4	-0.5	The student's reflection does not meet the criteria set. It is based on frustration rather than reflection, and not aligned with guiding questions. The language used is inappropriate, it tends to lack respect and attention to provide feedback.

5. Findings

This section reports how GPT-4 generated formative assessment within structured peer review exercises. The findings focus on the characteristics of the feedback produced, its alignment with assessment criteria, its structure features, and its comparison with human

marking practices. All observations are based on comparative analysis of AI-generated and human-generated feedback under shared rubric conditions.

5.1 Characteristics of GPT-4-Generated Formative Feedback

GPT-4 was configured using prompts aligned with learning objectives and the assessment rubric (see Methodology). The prompt design drew on the C.R.E.A.T.E framework (Birss, 2023) and the C.L.E.A.R. framework (Lo, 2023). Under these conditions, GPT-4 generated feedback that was timely, criterion-refenced, and structured. This configuration provide a consistent basis for examining how the AI-generated feedback related to task requirements and marking outcomes.

5.1.1 Alignment with Learning Objectives

The prompts used in this study were explicitly mapped to requirements of the peer review tasks, including peer commentary and reflective responses. As a result, AI-generated feedback consistently referenced relevant task requirements and rubric elements. The feedback addressed student performance in relation to clearly defined learning goals.

5.1.2 Timely and Relevant Feedback

, GPT-4's identification of full-mark cases and cases where students failed to attempt tasks closely aligned with the decisions made by human markers. As shown in Table 1 and Table 2, GPT-4 generated specific explanatory comments alongside assigned scores. In addition to content-specific feedback, GPT-4 identified structural features of submissions, such as instances where no peer review was submitted, and responded appropriately. The behavior reflects consistent application of feedback rules across submissions. However, the present study does not examine whether students perceive this feedback as helpful or how they apply it in their learning.

5.2 Feedback Structure Analysis Using Hattie and Timperley's Model

Following Hattie and Timperley's (2007) the analysis focused on the three guiding questions: “Where am I going?”, “How am I going?”, and “Where to next?” (Figure 3). This model was used descriptively to characterize the focus and organization of the feedback, rather than to evaluate feedback effectiveness or learning outcomes. GPT-4 was able to provide structured feedback that aligned with assignment criteria so students could understand task objectives as well as understand their performance. GPT-4 generated forward-looking comments that suggested areas for improvement

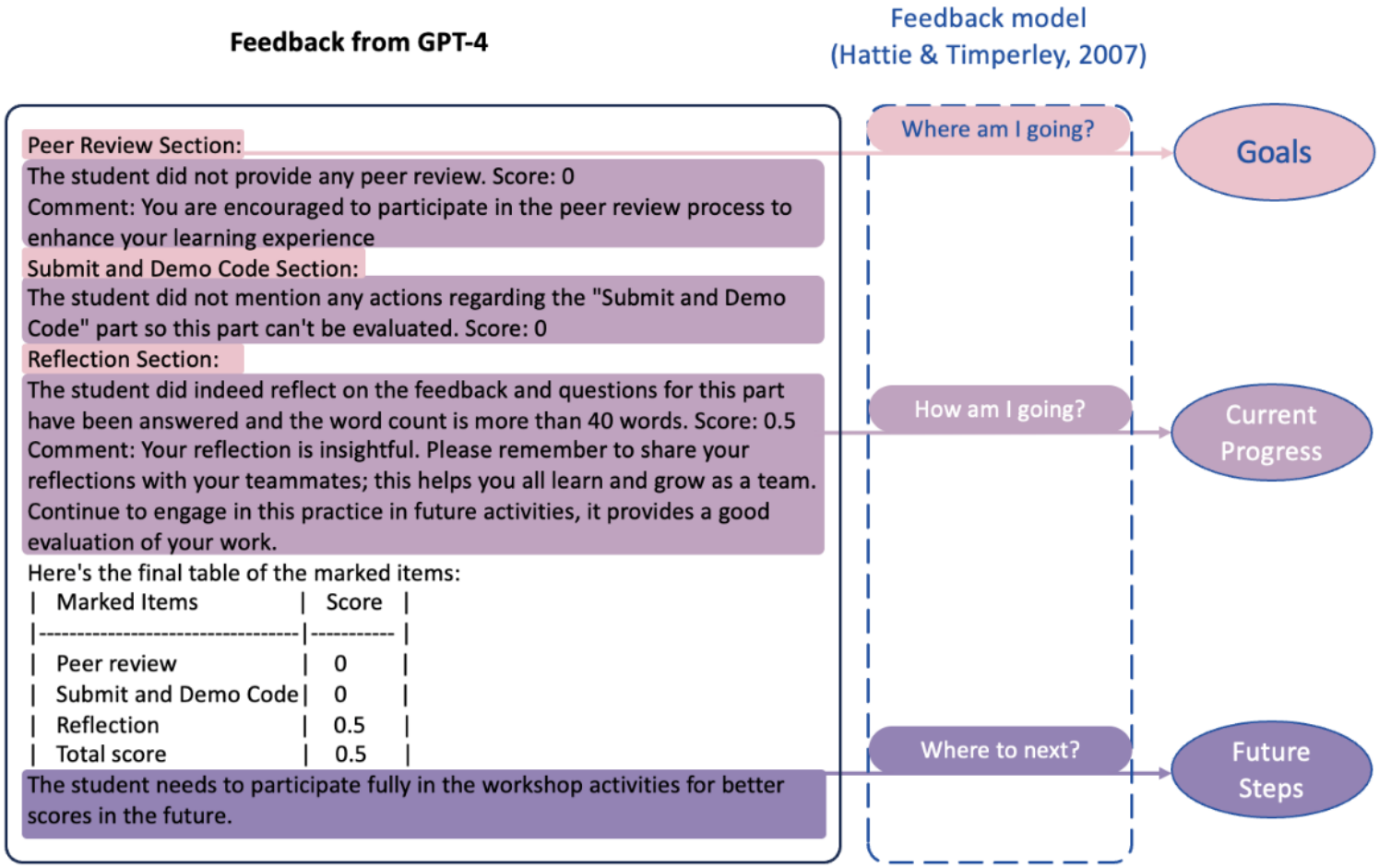

Figure 3 Analysis of GPT-4 Sample Feedback Using Hattie and Timperley's (2007) Feedback Model

5.3 Comparison of Feedback and Marking: GPT-4 vs. Human Markers

The comparative analysis examined both the feedback content and scoring alignment between GPT-4 and human markers. The results indicated that GPT-4 consistently generated more detailed feedback, providing specific comments related to each assessment criterion. For instance, in the sample feedback presented in Table 1, a human marker awarded a score of 2 without additional written comments, likely due to time limitations caused by the large volume of assessments. In contrast, GPT-4 assigned the same score while producing detailed, criterion-referenced explanations for each assessment component. GPT-4 identified specific strengths aligned with the rubric even in full mark submissions. Human markers, by comparison, tended to provide written feedback primarily when marks were deducted.

Regarding scoring outcomes, the analysis showed a high degree of agreement between GPT-4 and human markers in straightforward cases, including “Full Mark Cases” and “Deductions for Not Attempting Tasks”. To support interpretation of these findings, we categorize feedback tasks as either objective or subjective. Objective tasks followed clear, rule-based criteria, such as whether a student attempted a task, responded to guiding questions, or met the required word count. These tasks could be evaluated based on the presence or absence of specific features.

In contrast, subjective tasks required interpretive judgment, such as assessing the quality of peer feedback and the depth of student reflection. These judgments involve contextual considerations that are less explicitly specified in rubric descriptors. Differences between GPT-4 and human markers were most evident in these subjective areas. Human markers more frequently deducted marks for peer reviews that were brief or lacked specificity. GPT-4, however, tended to award full marks when submissions satisfied structural rubric requirements. These findings indicate that GPT-4 closely aligned with human judgment for

objective assessment dimensions but did not consistently reproduce the qualitative judgements applied by human educators.

6. Discussion

In this study, an LLM was able to analyze individual answers and provide explanations for why certain responses did not meet assessment criteria, guiding students towards recognizing errors and gaps in understanding. Such feedback exhibits features commonly associated with one-to-one tutoring, including immediacy and task-specific guidance (Bloom, 1984). When delivered feedback during the learning process, AI-generated feedback may help students identify and address knowledge gaps more promptly, supporting more accurate self-assessment and potential mitigating cognitive biases such as the Kruger-Dunning effect (Kruger & Dunning, 1999).

However, the findings also indicated limitations in the capacity of LLM-based feedback systems to replicate the nuanced judgments made by experienced educators. While the LLM aligned closely with human markers in straightforward assessment scenarios such as “Full Mark” and “Failure to Attempt Tasks”, its performance was less reliable in qualitative areas, including the quality of peer review and the depth of student reflection as noted by Dai et. al.(2023). Teachers can interpret nuances in peer reviews and the constructiveness of feedback through direct interaction with students, LLM-based feedback systems may struggle to capture these nuances, leading to differences in grading. These findings align with prior large-scale evidence showing that, although students value AI-generated feedback for its accessibility and clarity, teacher feedback continues to be regarded as more trustworthy, indicating the importance of human calibration in formative assessment settings (Henderson et al., 2025).

LLM-based feedback systems demonstrate capabilities that complement, and in some limited contexts exceed, traditional assessment practices for example when evaluating word count

maximums and minimums in assessment (Table 3: Reflection 1). In contrast to human graders, whose judgment may be affected by fatigue or inconsistency over time, LLM-based systems apply scoring criteria with consistency. Prior research has shown that fatigue can be impair human cognitive performance, and studies on AI-assisted grading suggest that automated systems may help reduce fatigue-related variability in large classroom settings (Gobrecht et al., 2024; Parekh et al., 2020).

These findings indicate that AI-generate feedback can be structured, detailed, and aligned with an established feedback framework (Hattie & Timperley, 2007). This structural alignment suggests potential value for supporting formative assessment at scale. However, whether such structural quality contributes to improved learning outcomes remains an open question. As noted by Carless (2022), the effectiveness of feedback depends not only on its quality but also on students' capacity to actively interpret, evaluate, and act on feedback within supportive learning environments. LLMs such as GPT-4 may play a supportive role in delivering formative feedback at scale, functioning as a learning resource rather than a determinant of learning impact. Therefore, further empirical investigation is needed to examine the effects of AI-generated feedback on student uptake and academic development.

In this study, human grading often provided written feedback primarily when marks were deducted, which may reduce the frequency and scope of constructive feedback available to students. In contrast, the LLM consistently applied structural criteria to student work and was able to address gaps often overlooked in routine feedback practices. This ability is particularly in large classes, where maintaining consistency and providing individualized attention present ongoing challenges.

The assessment design of the course further supports the use of LLM-generated in this setting. Each formative task carried a low weighting, minimizing the impact of discrepancies between AI-generated and human scoring. Only the best five out of seven peer reviews, each

contributing 2% to the total grade, were counted. This grading strategy is designed not as strict evaluation measures but as a motivational tool to encourage student engagement and effort. In practice, teachers often use point deductions to motivate students rather than penalize them. Within such low-stakes formative settings, the use of LLM-based feedback systems appears pedagogically justifiable, as minor scoring differences have limited consequences for students' overall performance.

From a sociocultural perspective, learning is mediated through interaction with others and with tools, and feedback can be understood as a form of social mediation that helps learners engage with standards and expectations (Vygotsky, 1978). LLMs may function as mediating artefacts embedded within peer review activities, supporting feedback process without replacing the dialogic and relational role of teacher-student interaction. Similarly, Zimmerman's (2008) self-regulated learning framework emphasizes the importance of feedback in enabling learners to monitor, evaluate, and adjust their learning strategies. While AI-generated feedback can contribute timely and structured input to support self-monitoring and goal setting, human involvement remains essential for providing adaptive, contextualized guidance and for supporting deeper reflective engagement.

Consistent with broader critiques of feedback practice in higher education, the educational value of AI-generated feedback depends not only on its level of detail but also on how it is embedded within learning designs that promote feedback use, dialogue, and reflection (Henderson et al., 2019). LLM-based feedback systems are most appropriately used when integrated into structured formative assessment designs that balance scalability and consistency with human judgment and pedagogical intent.

In this assessment setting, the use of an LLM supports the provision of continuous and structured formative feedback while helping to manage teaching workload in large classes. It shows the potential of LLM-based feedback systems to supplement traditional educational

practices, particularly where the focus is on developmental feedback rather than critical evaluation. Moving forward, a balanced human-AI approach that uses the scalability and consistency of AI with the essential role of human judgment in qualitative and contextual evaluation provides a promising and pedagogically sound direction for future assessment innovation.

7. Limitations and Directions for Future Research

This study is a preliminary, exploration examination of the use of a LLM implemented here through GPT-4, as a source of rubric-aligned formative feedback within a single course setting. While the study attempted to cover a range of assessment scenarios, several limitations remain. These include a small sample size, a potential imbalance in the distribution of submission categories, reliance on a single course context, and the absence of student perspectives. To better examine the applicability of these approaches, future research should extend this work to additional courses and live teaching settings.

In addition to the constraints already noted, the relatively small analytic sample drawn from a single course context limits the statistical power and generalizability of the findings. Patterns of alignment between GPT-4 and human marking may be influenced by instructor practices, rubric design, and cohort-specific factors that are not representative of other courses or disciplines. Moreover, the dataset included two simulated (constructed) edge-case submissions created to probe model behavior; while useful for testing responses to atypical situations, these simulated cases do not fully capture the complexity, variability, and contextual nuance of authentic student work, thereby reducing ecological validity. Together, these factors indicate that the results should be treated as exploratory and illustrative rather than broadly generalizable; replication with larger, more diverse samples and fully authentic submissions—ideally within live course implementations that examine student uptake and learning outcomes—will be necessary to establish wider applicability.

Based on the findings of this exploratory study, future research should focus on several key areas to address these limitations and better understand the role of AI in education:

Understand Student Use of Feedback: The next phase of our research will explore how students engage with feedback from both teachers and AI systems. Using observations, focus groups, surveys, and performance data, we aim to understand how students integrate this feedback into their learning. The findings will provide valuable insights into the impact of AI-assisted formative assessment on learning engagement, and perceptions of feedback quality in practical education settings.

Enhancing AI Capabilities: Fine-tuning LLMs for specific educational environments, as suggested by Y. Dai et al. (2023), can significantly improve their performance. This involves adapting models to better understand and respond to the nuanced communication and diverse educational needs across different learning environments. In addition, future studies could explore the "Extras" factor in the C.R.E.A.T.E. framework (Birss, 2023), which was mentioned in the Methodology section but was not used in this study, may improve prompt design and help align AI-generated feedback with pedagogical goals.

Developing AI-Human Collaborative Models: Investigating effective AI-human collaborative models remains a priority. As described by Molenaar (2022), AI can support educators by managing routine or clearly defined feedback tasks, while human teachers focus on complex evaluative judgments and relational aspects of teaching. Further research is needed to understand how these collaborative approaches can support formative assessment without reducing pedagogical quality.

Live Course Integration: Extending this research into live course settings is important for understanding how AI-generated feedback works under real teaching conditions. Studying AI-supported feedback in active classrooms can provide deeper insights into how it affects feedback processes, student engagement, and learning outcomes over time.

Educators' Perspectives and Collaboration: Including educators' perspectives on AI capabilities and their firsthand experiences with these tools in educational practice is critical. Collaborative research involving educators may help ensure that LLM-based feedback systems are not only technologically effective but also practical and meaningful in everyday teaching.

Ethical Concerns in AI Integration: Ongoing monitoring is needed to identify and reduce potential biases that may disadvantage some student groups. Protecting student privacy and meeting data protection requirements are also essential. In addition, the integration of AI may change the traditional student-teacher relationship and risk widening the digital divide. These concerns call for the development of clear guidelines and safeguards are needed to support fair and responsible use.

Longitudinal Studies on AI Impact: Long-term studies are needed to understand the impact of AI use across different educational settings and student groups. Such studies can help examine how AI affects equity and teaching quality over time.

Exploring Gender Dimensions: This study did not collect gender-specific data, but prior research suggests that gender may influence how students engage with AI-based tools. For example, Ofosu-Ampong (2023), Møgelvang et al. (2024). These findings suggest the need for gender focused analysis in future studies, especially in STEM fields where different student groups may experience AI-generated feedback in different ways.

Despite limitations of the study, useful insights were yielded and this initial study forms a strong foundation for the next phases of this research. Above all, the study reinforces that AI-generated feedback works best as a supporting tool. Educators should not view it as a replacement for professional judgment. Effective use depends on assessment design. Clear criteria, structured prompts, and ongoing human oversight are necessary. When these

conditions are met, AI-based feedback can support formative assessment practice in a variety of contexts.

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available as the data consists of students' assessment submissions, but are available from the corresponding author on reasonable request.

Acknowledgements

We would like to thank the students and teachers from the Web Application Development course for their participation and invaluable contributions to this study.

Declaration of the Use of Generative AI

During the preparation of this work, the authors used Open AI's Chat GPT for language editing and formatting. Following the use of this tool, the authors reviewed and revised the content as necessary, assuming full responsibility for the published content.

References

Abdel Aziz, M. H., Rowe, C., Southwood, R., Nogid, A., Berman, S., & Gustafson, K. (2024). A scoping review of artificial intelligence within pharmacy education. *American Journal of Pharmaceutical Education, 88*(1), 100615. https://doi.org/10.1016/j.ajpe.2023.100615

Ali, K., Barhom, N., Tamimi, F., & Duggal, M. (2023). ChatGPT—A double-edged sword for healthcare education? Implications for assessments of dental students. *European Journal of Dental Education*. Advance online publication. https://doi.org/10.1111/eje.12937

Ballantine, J., Boyce, G., & Stoner, G. (2024). A critical review of AI in accounting education: Threat and opportunity. *Critical Perspectives on Accounting, 99*, 102711. https://doi.org/10.1016/j.cpa.2024.102711

Beerepoot, M. T. P. (2023). Formative and summative automated assessment with multiple-choice question banks. *Journal of Chemical Education, 100*(8), 2947–2955. https://doi.org/10.1021/acs.jchemed.3c00120

Birss, D. (2023). *The prompt collection*. (Publisher not provided.)

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

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

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education, 38*(6), 698–712. https://doi.org/10.1080/02602938.2012.691462

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. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Carless, D. (2022). From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes. *Active Learning in Higher Education, 23*(2), 143–153. https://doi.org/10.1177/1469787420945845

Chang, D. H., Lin, M. P.-C., Hajian, S., & Wang, Q. Q. (2023). Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. *Sustainability, 15*(17), 12921. https://doi.org/10.3390/su151712921

Correia, A.-P., Hickey, S., & Xu, F. (2025). Realizing the possibilities of the large language models: Strategies for prompt engineering in educational inquiries. *Theory Into Practice, 64*(4), 434–447. https://doi.org/10.1080/00405841.2025.2528545

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 *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 323–325). IEEE. https://doi.org/10.1109/ICALT58122.2023.00100

Dai, Y., Liu, A., & Lim, C. P. (2023). Reconceptualizing ChatGPT and generative AI as a student-driven innovation in higher education. *Procedia CIRP, 119*, 84–90. https://doi.org/10.1016/j.procir.2023.05.002

Gao, R., Merzdorf, H. E., Anwar, S., Hipwell, M. C., & Srinivasa, A. R. (2024). Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence, 6*, 100206. https://doi.org/10.1016/j.caeai.2024.100206

Gobrecht, A., Tuma, F., Möller, M., Zöller, T., Zakhvatkin, M., Wuttig, A., Sommerfeldt, H., & Schütt, S. (2024). Beyond human subjectivity and error: A novel AI grading system (arXiv:2405.04323). *arXiv*. https://doi.org/10.48550/arXiv.2405.04323

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Henderson, M., Bearman, M., Chung, J., Fawns, T., Buckingham Shum, S., Matthews, K. E., & De Mello Heredia, J. (2025). Comparing generative AI and teacher feedback: Student perceptions of usefulness and trustworthiness. *Assessment & Evaluation in Higher Education*, 1–16. https://doi.org/10.1080/02602938.2025.2502582

Henderson, M., Ryan, T., & Phillips, M. (2019). The challenges of feedback in higher education. *Assessment & Evaluation in Higher Education, 44*(8), 1237–1252. https://doi.org/10.1080/02602938.2019.1599815

Irons, A., & Elkington, S. (2021). *Enhancing learning through formative assessment and feedback* (2nd ed.). Routledge.

Kerman, N. T., Banihashem, S. K., Karami, M., Er, E., Van Ginkel, S., & Noroozi, O. (2024). Online peer feedback in higher education: A synthesis of the literature. *Education and Information Technologies, 29*(1), 763–813. https://doi.org/10.1007/s10639-023-12273-8

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

Lee, D., & Palmer, E. (2025). Prompt engineering in higher education: A systematic review to help inform curricula. *International Journal of Educational Technology in Higher Education, 22*(1), 7. https://doi.org/10.1186/s41239-025-00503-7

Lee, Y.-C., & Fu, W.-T. (2019). Supporting peer assessment in education with conversational agents. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion* (pp. 7–8). https://doi.org/10.1145/3308557.3308695

Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship, 49*(4), 102720. https://doi.org/10.1016/j.acalib.2023.102720

Memarian, B., & Doleck, T. (2024). A review of assessment for learning with artificial intelligence. *Computers in Human Behavior: Artificial Humans, 2*(1), 100040. https://doi.org/10.1016/j.chbah.2023.100040

Messer, M., Brown, N. C. C., Kölling, M., & Shi, M. (2024). Automated grading and feedback tools for programming education: A systematic review. *ACM Transactions on Computing Education, 24*(1), 1–43. https://doi.org/10.1145/3636515

Møgelvang, A., Bjelland, C., Grassini, S., & Ludvigsen, K. (2024). Gender differences in the use of generative artificial intelligence chatbots in higher education: Characteristics and consequences. *Education Sciences, 14*(12), 1363. https://doi.org/10.3390/educsci14121363

Molenaar, I. (2022). The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning. *Computers and Education: Artificial Intelligence, 3*, 100070. https://doi.org/10.1016/j.caeai.2022.100070

Ng, S. W. (2012). The impact of peer assessment and feedback strategy in learning computer programming in higher education. *Issues in Informing Science and Information Technology, 9*, 17–27. https://doi.org/10.28945/1601

Nicol, D. D., & Macfarlane-Dick, D. (2006). Rethinking formative assessment in higher education: A theoretical model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218.

Ocampo, J. C. G., & Panadero, E. (2023). Web-based peer assessment platforms: What educational features influence learning, feedback and social interaction? In O. Noroozi & B. De Wever (Eds.), *The power of peer learning* (pp. 165–182). *Springer International Publishing*. https://doi.org/10.1007/978-3-031-29411-2_8

Ofosu-Ampong, K. (2023). Gender differences in perception of artificial intelligence-based tools. *Journal of Digital Art & Humanities, 4*(2), 52–56. https://doi.org/10.33847/2712-8149.4.2_6

Parekh, V., Shah, D., & Shah, M. (2020). Fatigue detection using artificial intelligence framework. *Augmented Human Research, 5*(1), 5. https://doi.org/10.1007/s41133-019-0023-4

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–7). https://doi.org/10.1145/3411763.3451760

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wong, J., Baars, M., Davis, D., Van Der Zee, T., Houben, G.-J., & Paas, F. (2019). Supporting self-regulated learning in online learning environments and MOOCs: A systematic review. *International Journal of Human–Computer Interaction, 35*(4–5), 356–373. https://doi.org/10.1080/10447318.2018.1543084

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal, 45*(1), 166–183. https://doi.org/10.3102/0002831207312909