ASSESSMENT & EVALUATION IN HIGHER EDUCATION
https://doi.org/10.1080/02602938.2024.2415649

Routledge
Taylor & Francis Group

Exploring the use of artificial intelligence (AI) in the delivery of effective feedback

Julia Venter, Stephen A. Coetzee and Astrid Schmulian

Department of Accounting, University of Pretoria, Pretoria, South Africa

ABSTRACT

Providing effective feedback in large settings presents significant challenges due to time and resource constraints. Given the importance of feedback in competency-based higher education, innovative solutions are essential. This study explores the integration of artificial intelligence (AI) to enhance feedback delivery. The research focuses on the development of a custom prompt for GPT-4 within a no-code web application, designed to deliver AI-generated feedback to second-year accounting students on discussion or essay-style questions in a large competency-based intermediate accounting course in South Africa. The prompt aligns with the principles of effective feedback and was tested in a pilot evaluation. Results indicated that the AI-generated feedback generally adhered to these principles, though some variability was observed across the different feedback dimensions. While challenges remain in consistently guiding AI to apply pedagogical best practices, the findings suggest that large language models can complement traditional feedback. Rigorous oversight of AI-generated feedback remains critical.

KEYWORDS

Artificial intelligence; automated feedback; ChatGPT; feedback

Introduction

Feedback is one of the most powerful means to support learning (Hattie 2008). However, delivering feedback effectively, especially in large educational settings, is challenging due to time and resource constraints (Boud and Molloy 2013; Cavalcanti et al. 2019; Dai et al. 2023; Demszky et al. 2023; Pardo et al. 2019). The critical role of effective feedback in enhancing learning, particularly in competency-based education, underscores the urgent need for innovative solutions (Hattie and Timperley 2007; Henderson et al. 2019; Nicol and Macfarlane-Dick 2006; Parikh, McReelis, and Hodges 2001; Tekian et al. 2017). Artificial intelligence's (AI) integration into feedback mechanisms, notably through advancements in natural language processing (NLP), presents a transformative opportunity to address these challenges (Bauer et al. 2023; Deeva et al. 2021; Grassini 2023; Hooda et al. 2022).

Owing to its focus on linguistic quality and performance analysis, NLP contrasts with the data-oriented approaches of educational data mining and learning analytics (Gardner, O'Leary, and Yuan 2021; Zhang et al. 2019). NLPs can process and generate feedback at scale, and can complement human instructors. The evolution of NLP, particularly through large language models (LLMs), such as the generative pre-trained transformer (GPT) models developed by OpenAI, marks a significant advancement in language processing capabilities. LLM's ability to process and

CONTACT Stephen A. Coetzee stephen.coetzee@up.ac.za University of Pretoria, South Africa

generate feedback at scale can ensure that students receive feedback that is timely, personalised and at low cost to instructors (Kasneci et al. 2023). These advantages potentially make LLMs a suitable tool for enhancing textual feedback processes (Kung et al. 2023), particularly in large-scale educational contexts, while freeing up time for instructors to create more innovative lesson plans, engage in professional development, promote research and support students (Alshater 2022; Grassini 2023; Terwiesch 2023), all of which are influential in enhancing students' learning (Sok and Heng 2023).

Despite the potential of LLMs to provide feedback, there is some concern around their utility (Baidoo-Anu and Owusu Ansah 2023; Chomsky, Roberts, and Watumull 2023; Grassini 2023; Pegoraro et al. 2023). LLMs create responses by predicting the likelihood of word sequences based on large amounts of data (Bender et al. 2021; Marcus and Davis 2020). While this allows them to produce text that is intelligible, their outputs are based on statistical patterns rather than a true understanding of meaning or context. This limitation can lead to responses that, despite being linguistically sound, may lack depth, logical coherence or relevance. The integration of LLMs into educational feedback raises important ethical considerations, particularly the potential for bias (Baidoo-Anu and Owusu Ansah 2023; Grassini 2023). AI systems are trained on large datasets that may contain inherent biases, which could manifest in the feedback provided to students. However, one of the ways in which these risks can be mitigated, in conjunction with instructor oversight, is through prompt engineering – the practice of formulating clear and specific instructions that guide the LLM towards a desired output (Schramowski et al. 2022; White et al. 2023).

This paper outlines the development of a prompt aimed at delivering AI generated feedback that aligns with established principles of effective feedback to second year students in a large, competency-based accounting course at a South African university on their discussion or essay-style questions. The development of the prompt was guided by Nicol and Macfarlane-Dick's (2006) seminal framework on effective feedback. The prompt was crafted within a self-developed web application, created using the no-code platform Bubble.io, that facilitates the submission of students' discussion or essay-style questions and delivers corresponding feedback from GPT-4.

This paper contributes to the feedback and pedagogical technology literature by demonstrating the development of the prompt for use with GPT-4, within a self-developed web application, crafted with Bubble.io, to automate feedback generation. The choice of Bubble.io reflects its accessibility to instructors without coding expertise, potentially broadening the application's use across various educational contexts. By exploring the use of AI-generated feedback within an undergraduate accounting course, this paper aims to underscore the potential of AI to contribute towards feedback effectiveness in large learning environments.

Theoretical background

Effective feedback

The foundation of this study is rooted in well-established educational and technological theories that support the integration of AI into feedback mechanisms. The study draws upon Nicol and Macfarlane-Dick's (2006) framework of effective feedback, which outlines seven principles that are critical for fostering student learning through feedback. These principles guide AI-generated feedback, ensuring that the feedback provided by AI aligns with proven educational strategies.

Effective feedback plays a pivotal role in education, serving as a critical tool through which students assimilate various sources of information to enhance their learning and performance (Carless and Boud 2018; Henderson et al. 2019). Effective feedback bridges the gap between current achievements and aspirational goals, acting as a catalyst for academic and personal growth (Hattie and Timperley 2007). While effective feedback has the potential to significantly boost learning outcomes (Nicol and Macfarlane-Dick 2006; Parikh, McReelis, and Hodges 2001),

insufficient or misaligned feedback can detrimentally impact student motivation and progress (Cavalcanti et al. 2019).

Nicol and Macfarlane-Dick (2006) underscored the essence of effective feedback by delineating seven fundamental principles of effective feedback. These principles serve as a comprehensive guide for instructors, aiming to refine feedback practices in a manner that elevates student learning. These principles have been extensively validated in various educational contexts, demonstrating their effectiveness in enhancing student learning and engagement. They have also been empirically supported by numerous studies, establishing their robustness as a foundation for developing feedback mechanisms, whether human or AI-generated.

Principles of effective feedback

Clarity of what good performance is (i.e. the goal). Student engagement with learning tasks increases significantly when clear, specific criteria for success are explicitly communicated (Fisher and Frey 2009; Hattie and Timperley 2007; Nicol and Macfarlane-Dick 2006). Clarity in the goals of learning tasks ensures that feedback is more impactful, as it aligns with well-defined targets that students are aware they should meet. This alignment not only enhances the effectiveness of feedback by providing a clear benchmark for success but also increases motivation and minimizes student frustration by reducing ambiguity about what constitutes satisfactory performance (Balloo et al. 2018). By understanding the goals and standards they are expected to meet, their ability to self-regulate their learning also improves. A proven strategy for clarifying these goals involves providing students with exemplars of expected performance (Orsmond, Merry, and Reiling 2002). Exemplars serve as concrete benchmarks, illustrating the required standards and quality of work. By engaging in comparative analysis between their work and the exemplars, students can gain a deeper understanding of their current standing relative to the goals of the learning task and identify specific areas for improvement.

Opportunities to facilitate self-assessment. Facilitating self-assessment is pivotal in enabling students to critically evaluate their own learning and work. The importance of self-assessment in feedback processes has been underscored by Carless and Boud (2018), who demonstrated that encouraging students to evaluate their own work fosters deeper learning and reflective practices. By actively engaging in self-assessment, students can identify areas of strength and those requiring improvement, fostering a deeper understanding of their learning journey (Nicol and Macfarlane-Dick 2006). This reflective practice empowers students to internalize feedback and take ownership of their academic progress, enhancing their ability to apply feedback constructively to guide their future learning.

To support self-assessment, instructors can integrate various tasks and activities into the curriculum that encourage students to reflect on their learning and solicit specific feedback. This might involve students specifying the type of feedback they require upon submission or self-evaluating their work against given exemplars before submission. Examples of such activities include reflective journals, which allow for ongoing self-assessment, and peer assessments, where students can compare their work with that of their classmates. The use of exemplars provides clear benchmarks for performance, enabling students to articulate their feedback needs more effectively and engage in targeted self-review (Boud, Keogh, and Walker 2013).

Within this framework of self-assessment, developing feedback literacy emerges as a key component. Feedback literacy equips students with the skills to actively seek, understand and apply feedback, and to contribute to the feedback process, thereby enhancing their academic development (Carless and Boud 2018; Nicol and Macfarlane-Dick 2006). The instructors' role is critical in nurturing this literacy, guiding students through the process of engaging with feedback, recognizing the value of constructive criticism, and implementing feedback for improvement.

Delivery of high-quality information to students about their learning. The essence of effective feedback lies in its ability to provide detailed, goal-aligned insights that are both understandable and actionable for students (Boud and Molloy 2013; Cavalcanti et al. 2020; Nicol and Macfarlane-Dick 2006), that they can use to improve their performance. High-quality feedback transcends generic praise or criticism, offering clear guidance tailored to the objectives of the assignment and the individual needs of the student. This detailed feedback, ideally delivered by instructors, serves as a crucial benchmark, aiding students in calibrating their self-assessment and understanding the nuances of their performance (Higgins, Hartley, and Skelton 2001).

The timing of feedback delivery is equally pivotal, necessitating a balance that considers the instructional cycle's dynamics (Boud and Molloy 2013; Hattie and Timperley 2007; Yang and Carless 2013). Immediate feedback on discrete concepts can prompt timely adjustments in learning strategies, while delayed feedback for broader projects can incentivize self-reflection and assessment, enriching the learning experience.

Encouragement of instructor and peer dialogue around learning. The effectiveness of feedback can be increased by conceptualizing feedback as a dialogue rather than merely as information transmission (Laurillard 2013; Nicol and Macfarlane-Dick 2006). Viewing feedback as a dialogue means that students not only receive initial feedback information but can also discuss that feedback with their instructors. This perspective transforms feedback from simply being comments provided by instructors about students' work, to a process that requires active and ongoing student engagement to promote learning (Boud and Molloy 2013; Henderson et al. 2019). Engaging in discussions with their instructors helps students develop their understanding of expectations and standards, clarify any misunderstandings, and receive immediate responses to any difficulties they may encounter, which is crucial for effective and deep learning (Winstone et al. 2017).

A helpful way to encourage dialogue is to ask students for examples of feedback that helped them improve and explain how it did so. To maintain the relational dimension of feedback, it is important that instructors are supportive, approachable and sensitive when providing feedback (Carless and Winstone 2023).

Encouragement of positive motivational beliefs and self-esteem. The emotional impact of feedback should not be overlooked, as emotions play a significant role in learning and assessment (Carless and Boud 2018; Nicol and Macfarlane-Dick 2006). While generic praise or criticism may be less useful to students (Henderson et al. 2019), it is essential to recognize that students' emotional reactions to feedback can influence their sense-making and motivation. Consequently, instructors should prioritize supportiveness, approachability and sensitivity when providing feedback (Carless and Winstone 2023).

Motivation and self-esteem are more likely to be enhanced when a course includes numerous low-stakes assessment tasks, with feedback aimed at providing information about progress and effort, focusing on how students can improve rather than merely on outcomes (Hattie and Timperley 2007). This approach is also supported by the work of Dweck (2000), who differentiates between a fixed and growth mindset. A fixed mindset is one in which students believe there is a limit to what they can achieve, while a growth mindset involves students believing that their ability is malleable and depends on the effort they put into a task. With a fixed mindset, students interpret failure as a reflection of their low ability and are likely to give up. In contrast, with a growth mindset, students view failure as a challenge or an obstacle to be overcome and increase their effort. Feedback focusing on progress and improvement encourages a growth mindset in students. By prioritising these factors when providing feedback, instructors can foster a conducive learning environment that encourages students to engage with feedback, learn from it, and strive for continuous improvement.

Opportunities to close the gap between current and desired performance. The idea that feedback should guide students on how to close the gap between their current and desired performance is central to effective learning, changes in student behaviour and improved performance (Yorke 2003). To ensure this outcome, actionable feedback, coupled with opportunities for students to apply it, leads to significant improvements in learning outcomes (Boud 2000; Evans 2013; Sadler 1989). Effective feedback design involves aligning multiple assessment tasks with linked competencies and interspersing them with opportunities for students to seek and receive information that can influence their subsequent tasks (Boud and Molloy 2013). This approach allows each feedback cycle to build on the previous one, creating a continuous feedback loop. Regular and varied feedback loops increase the likelihood of students understanding and acting upon crucial information, ultimately fostering a more effective learning environment. Feedback should be specific to provide students with actionable steps for improvement (Henderson et al. 2019).

Provision of information to instructors that can be used to help shape teaching. Effective feedback not only provides useful information that helps students improve their learning but also offers valuable insights to instructors to inform their instructional practices (Nicol and Macfarlane-Dick 2006). Feedback provided to students can serve as a diagnostic tool for instructors, helping them tailor their teaching strategies to better meet the needs of their students (Boud and Molloy 2013; van de Pol, Volman, and Beishuizen 2010; Yorke 2003). This reciprocal process ultimately enhances the learning experience for both students and instructors.

Effective feedback is a vital component of the learning process, as it helps students recognize and close the gap between their current performance and desired goals (Butler and Winne 1995). By adhering to the principles outlined, instructors can provide feedback that is clear, actionable and timely, ultimately fostering an environment that promotes continuous improvement.

Given the strong empirical support for these principles in human-generated feedback, the development of AI-generated feedback in this study was guided by the same principles to ensure its validity and effectiveness. This research aims to demonstrate that AI can effectively complement traditional feedback methods, particularly in large-scale educational contexts.

Integration of AI into feedback mechanisms

In large educational settings, providing high-quality feedback that is both timely and consistent remains a significant challenge, especially when balancing the need for personalized insights with the practical constraints of teaching at scale. AI, particularly NLP systems, offers a promising solution to this problem by automating the feedback process. Automated writing evaluation (AWE) systems, for instance, have been used to assess and provide feedback on essay-style or discussion-based questions. While effective in delivering quick feedback, these systems are limited by their task-specific design and high cost and development time for new applications (Nunes et al. 2022; Ramesh and Sanampudi 2022; Rupp et al. 2019).

Emerging advances in LLMs like GPT-4 present an alternative to traditional AWEs by offering more flexibility in handling diverse tasks without requiring extensive retraining. LLMs have demonstrated their ability to generate human-like feedback on a wide range of topics, including more open-ended, qualitative tasks (Pinto et al. 2023). This adaptability allows educators to provide tailored feedback on complex assignments, such as essays or project-based assessments, in real-time and at scale. Instructors can further refine the feedback process by using specific prompts to guide the AI's output, ensuring that feedback aligns with the pedagogical goals of the course (Kirk et al. 2022).

Prompt engineering

The task of formulating an optimal instruction or prompt for a LLM, often referred to as 'prompt engineering' (Bommarito et al. 2023), is integral to the successful application of LLMs. This task entails an iterative process of drafting, testing and refining to elicit desired responses. Given that LLMs are finely attuned to the nuances of input prompts (Kirk et al. 2022), prompt engineering must be approached with an understanding of the models' sensitivities. Additionally, it is imperative to acknowledge the inherent limitations and biases within LLMs, which can influence their response patterns (Borji 2023; Ray 2023). Examples of limitations and biases inherent in GPT-4's design include:

- **Outdated Knowledge**. The training data for the model used in the web application, developed in 2023, extended only up to April 2023, omitting any advancements that occurred thereafter. This limitation could potentially diminish the feedback's pertinence to current practices.
- **Ambiguity in Prompts**. Ambiguous responses from students can present difficulties for GPT-4, potentially leading to feedback that is less precise or helpful.
- **Over-Optimization**. The feedback, while coherent, may not specifically address the nuances of the student's submission, reflecting the model's training rather than the student's needs.
- **Inherent Biases**. Biases from the internet-based training corpus, such as those related to gender or race, can manifest in the feedback, at times leading to insensitivity.
- **Domain Expertise**. GPT-4's generalist training means it may lack the nuance that a domain expert would provide.
- **Repetition and Verbosity**. The model may produce feedback that is repetitive or overly verbose, which can dilute the clarity of the message.
- **Accuracy Issues**. On occasion, the feedback may be inaccurate or nonsensical, revealing limitations in the model's contextual understanding.

To mitigate these limitations and biases, it is essential to carefully design the prompt and iteratively refine it based on instructor evaluation and comparison.

Given the potential of LLMs, especially when effectively prompted, to produce feedback for discussion or essay style questions at scale, this paper extends the current body of research by applying the established framework of Nicol and Macfarlane-Dick's (2006) seven fundamental principles of effective feedback to the specific context of AI-generated feedback.

The study is guided by the following research question: to what extent does GPT-4, when prompted to deliver feedback aligned with established educational feedback principles, provide effective feedback to second year students in a large, competency-based accounting course on their discussion or essay-style questions?

Method

Developing an effective prompt for educational feedback

To inform the discussion on the development of the prompt, the authors, as 'complete participants' (Gold 1958) in the development process, relied upon their development notes, personal experiences, conversations and reflections during the period prior to and since the launch of the web application.

The journey from ChatGPT to customized application

In December 2022, following the launch of the ChatGPT prototype by OpenAI, the authors discussed the idea of using ChatGPT to provide feedback to students in a large accounting class. Some potential challenges were identified, including:

- Feedback generated by ChatGPT may not be contextually appropriate or fully aligned with the specific curriculum, learning objectives or assessment criteria set by instructors, potentially leading to confusion or misinterpretation by the students.
- Students would be learning outside the instructors' sphere of influence as instructors do not have any inputs in the feedback generated by ChatGPT.
- Instructors would not have access to or be able to review the feedback provided to students by ChatGPT.
- Students would be required to prompt ChatGPT themselves to obtain feedback on their work and they may not design the prompt optimally.
- Students might become distracted in their use of ChatGPT with other discussions with the LLM.
- Students may have unequal levels of access to ChatGPT's models. At the time of writing, ChatGPT offered a free version based on the GPT-3 model, while the more advanced GPT-4 model was available for paid subscribers.

The authors discussed developing their own dedicated web application to overcome these challenges. This application would be designed to provide effective feedback for second year accounting students on their discussion or essay-style questions. The web application would incorporate a prompt built around pre-set parameters, such as stating the discipline (accounting) and integrating the principles of good feedback (Nicol and Macfarlane-Dick 2006). The web application would guide students in inputting their questions, answers and suggested solution. It would also standardize the AI model used (GPT-4), mitigating issues related to prompt optimization and access inequality. The web application would prioritize interactivity, user-friendliness and the goal of enhancing the feedback experience. It would enable the instructors to see the feedback generated to their students.

Crafting the web application: a no-code approach

With minimal or no coding expertise, the authors chose Bubble.io as the tool to develop the envisaged web application. Bubble.io is a visual web development platform that allows users to build web applications without writing code. It provides a user-friendly drag-and-drop interface (Image 1), enabling non-technical users to create web applications by visually designing their user interface and configuring workflows to define how the application should behave.

The web application is structured to support several key functions in the feedback loop. It allows students to submit discussion or essay-style questions, input requests for targeted feedback on specific areas of their work needing guidance and input an optional self-reflection regarding the strengths and weaknesses of their answer, and where they struggled (Image 2). The system is also designed to facilitate an ongoing dialogue between the students and the LLM, enabling them to engage further with the feedback provided (Image 3).

The development of the web application involved an iterative process, with the authors rigorously testing each feature (such as the upload of students' answers) as it was developed to ensure functionality met expectations. The authors engaged in continuous refinement, aligning the design with pedagogical goals and user experience standards. This hands-on approach allowed for real-time adjustments and improvements, ensuring that each aspect of the application functioned effectively.

Integrating the prompt with GPT-4

In developing the web application, integrating it with OpenAI's GPT-4 *via* the OpenAI application programming interface (API) is a key feature. The API facilitates communication between the Bubble.

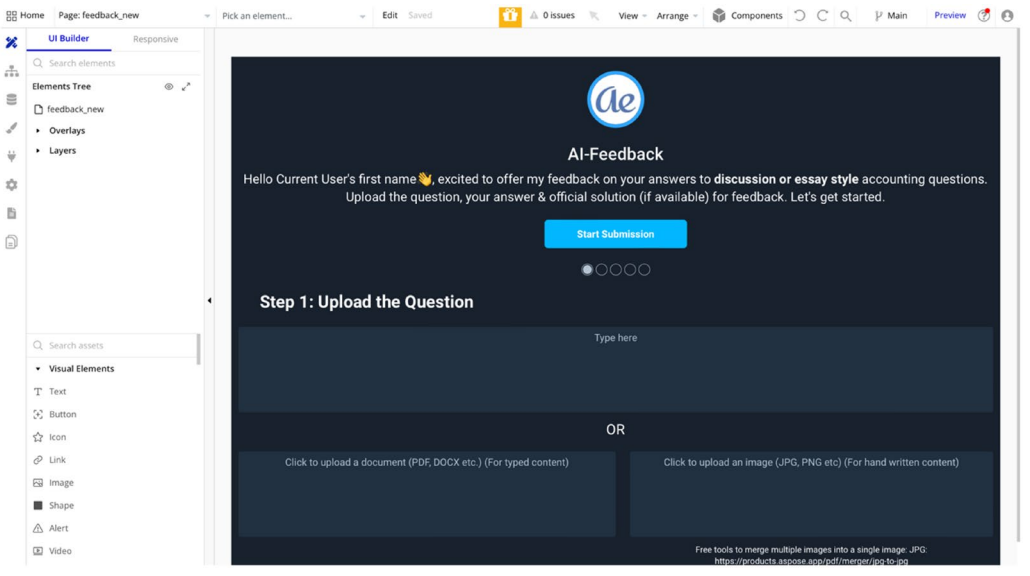

Image 1. User interface of the AI-generated feedback application developed on *Bubble.io*.

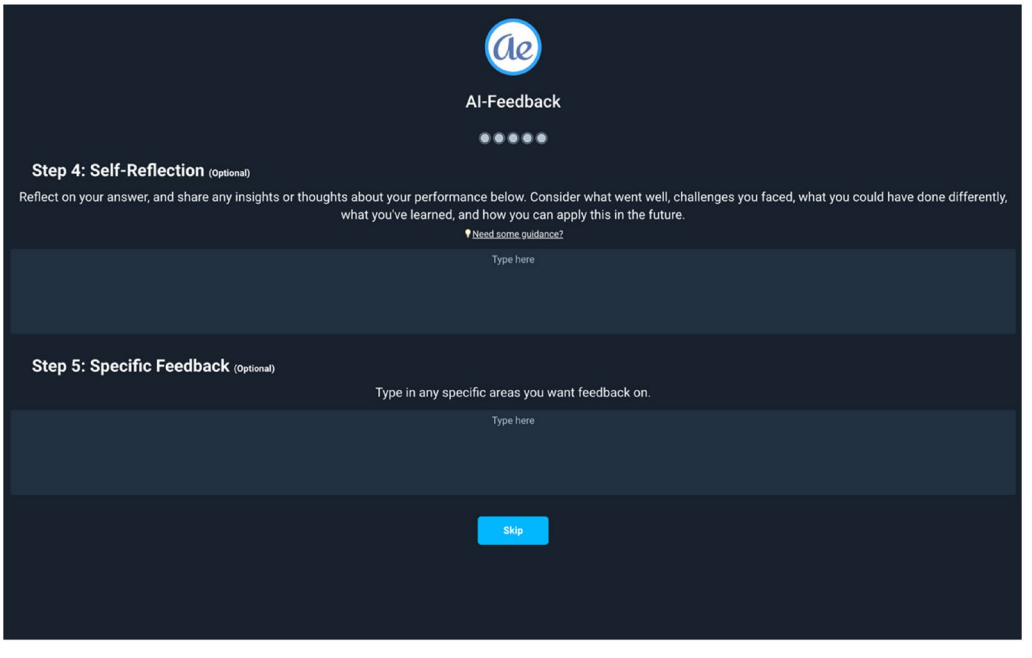

Image 2. Optional self-reflection regarding the strengths and weaknesses of their answer.

io-built web application and OpenAI, allowing for the submission of students' answers along with a carefully designed prompt. GPT-4 processes this information to generate and return tailored feedback, highlighting the critical role of prompt design in obtaining effective feedback from the LLM (Figure 1).

To ensure the privacy and security of data exchanges with OpenAI's API, all communications are encrypted. OpenAI enforces strict access controls, such as API key and OAuth2 authentication, and adheres to transparent data usage policies. Information sent to OpenAI is neither used for model training without explicit consent nor retained beyond 30 days. This allows for the

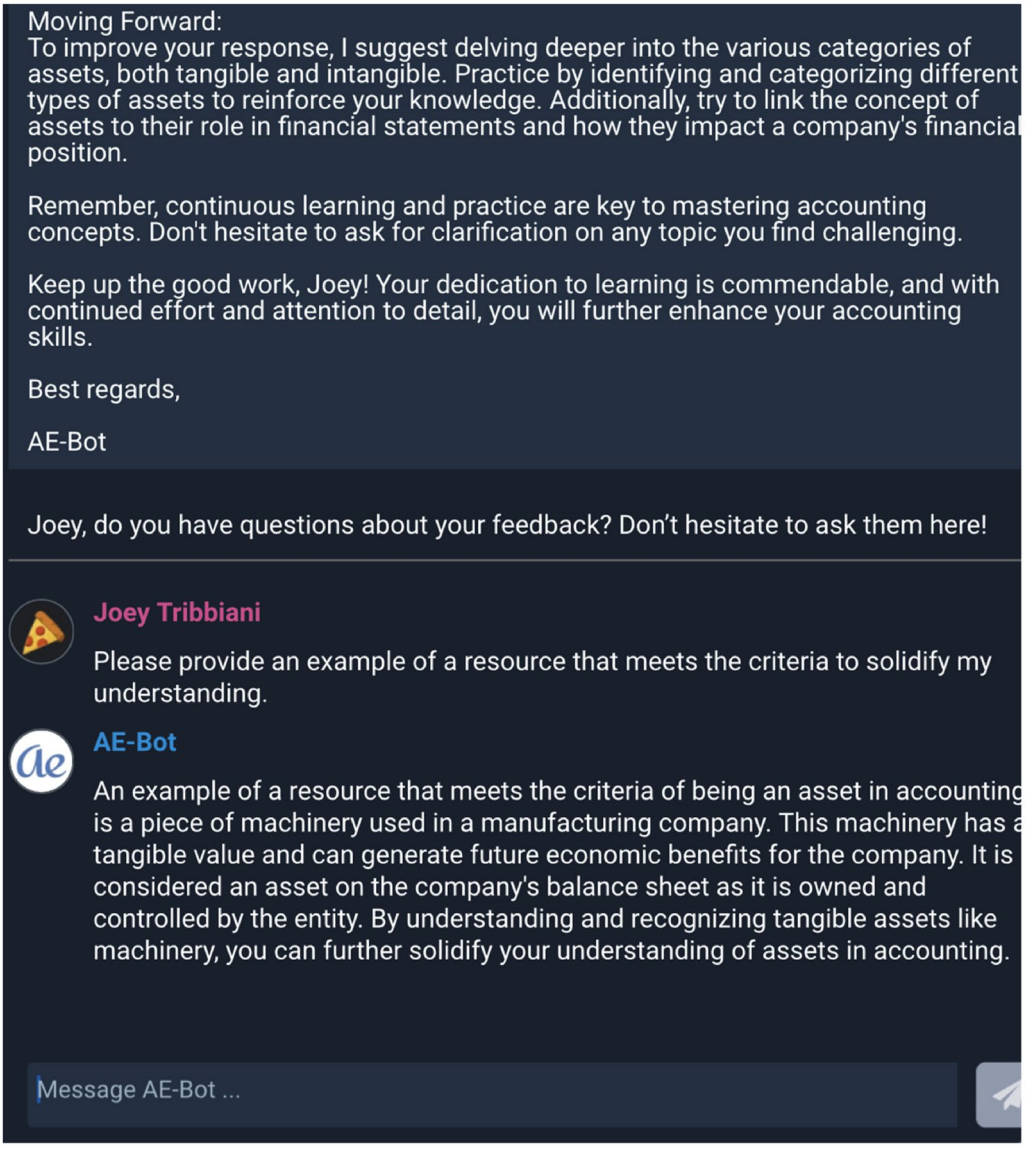

Image 3. AI-generated feedback and ongoing dialogue.

safeguarding of student data while utilizing the capabilities of OpenAI's LLM for educational purposes, distinct from the commercial ChatGPT product.

Drafting of the initial prompt

Within the context of LLMs, a prompt serves as the initial input – a query, statement or directive – that triggers the model's response (Beatman 2023). These prompts can range from broad, open-ended inquiries to targeted requests for specific information or action. The role of a prompt is to establish a frame of reference or an objective for the LLM, steering it toward generating outputs that are on-topic, applicable and informed by its vast training data (Borji 2023; White et al. 2023). A prompt acts not as a programming command, but rather as a guide; it does not compel a uniform response due to the stochastic nature of LLM operations. A well-crafted prompt

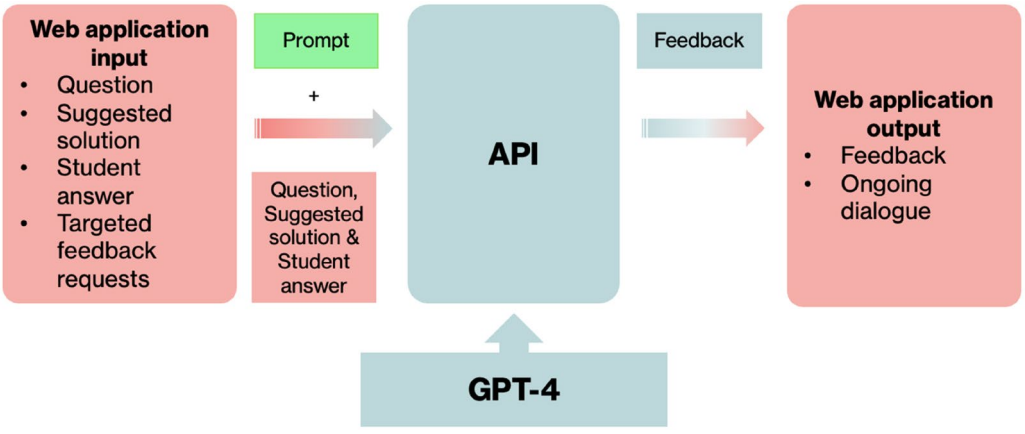

Figure 1. Process flow diagram of the AI-generated feedback system.

can significantly increase the likelihood of obtaining precise, pertinent and coherent responses, even with the inherent variability in LLM outputs (White et al. 2023).

The following template prompt obtained from the OpenAI API reference material (OpenAI n.d.) served as the point of departure for drafting the prompt:

```
{
  "model": "GPT-4",
  "temperature": 0.5,
  "messages": [{"role": "user", "content": "Hello!"}]
```

The initial prompt was written by one of the authors based on the principles of effective feedback proposed by Nicol and Macfarlane-Dick (2006). The prompt was then shared with the other authors, who reviewed the prompt and provided suggestions for improvement. Through discussions and consensus-building differences in opinion were resolved.

Evaluation of the effectiveness of the initial prompt. Following the initial drafting of the prompt an iterative refinement process, designed to enhance the quality of the feedback generated by the LLM, was conducted. This was done by analysing the feedback generated by the LLM in response to students' answer scripts of discussion or essay style questions.

Study setting, participants and task. The study's sample was drawn from the answer scripts of discussion or essay style questions from accounting students in their second year of study at a South African university. The answer scripts were drawn from five distinct accounting assessments over two academic years, each related to a different accounting case.

Data collection. After obtaining approval from the institutional review board (EMS064/23), a purposive sample of 15 student answer scripts from each assessment was selected. This strategy ensured a diversity in performance and coverage of accounting topics, cumulating in a total sample of 75 scripts. The consent of the students whose answer scripts were selected was sought. The students were advised that the study was anonymous and neither they, nor the content of their answer scripts, would be identifiable in any way, that they may withdraw their consent at any time throughout the study without negative consequences and that the results of the study would be used for academic purposes only. The adequacy of the sample size was affirmed by preliminary analysis, which indicated that it would be sufficient to reach data saturation – providing a comprehensive understanding of students' experiences and perspectives on feedback, without yielding redundant information.

Data analysis. Initially, a subset of these scripts ($n = 14$) from a single assessment was used to evaluate the prompt's effectiveness. These findings informed targeted revisions to the prompt's structure and content to align more closely with the feedback criteria. Subsequently, the finalized prompt was applied to the remaining 57 scripts (3 students opting out) across four additional assessments to confirm its effectiveness.

Following each revision, the modified prompt was redeployed to process the same set of answer scripts through the web application, enabling a comparative analysis of the feedback generated by the LLM. This iterative cycle of evaluation, refinement and re-evaluation was repeated until no substantial improvements in the quality of the feedback were observed, indicating that the prompt had reached its optimal form (Figure 2). This approach aimed to ensure that the feedback provided by the AI met a baseline standard of quality that could reasonably complement human-generated feedback.

A significant aspect of the refinement process involved adjustments to the sequence of instructions within the prompt. It was discovered that the order in which information was presented to the LLM significantly affected the focus and detail of the feedback. LLMs process information sequentially, meaning the arrangement of prompt components can guide the model's attention and response patterns. Strategic modifications to the structure of the prompt allowed the authors to direct the LLM's analysis towards the most relevant aspects of the student submissions, thereby enhancing the overall relevance and utility of the feedback provided.

Alignment between the feedback principles and the prompt

After finalizing the prompt, the alignment between the feedback principles derived from the literature and the specific elements of the prompt were tabularised (Table 1). The table provides a clear and organized representation of how the seven principles of effective feedback suggested by Nicol and Macfarlane-Dick (2006) were incorporated into the prompt.

Evaluation of the effectiveness of the final prompt

After refining the prompt based on insights from the initial evaluation of 14 scripts, the final version was applied to a broader dataset comprising the remaining 57 student scripts across four additional assessments. This expanded evaluation sought to confirm the effectiveness of the refined prompt, as evidenced by the improved quality and applicability of the feedback generated by the AI across a diverse range of student submissions. This evaluation was facilitated by a rubric (Table 2) developed by the authors to assess the feedback.

To ensure the rubric's reliability (Dawson 2017) – consistency of assessment across different applications – and its validity – accuracy in measuring what it is intended to measure – the authors grounded the rubric's design in the principles of effective feedback delineated by Nicol and Macfarlane-Dick (2006). This theoretical foundation provided a robust pedagogical basis for the rubric. Calibration sessions were conducted to align the authors interpretations of the rubric criteria.

These principles were subsequently transformed by the authors into specific, measurable criteria designed to assess the quality of AI-generated feedback accurately. To ensure the language of each criterion was clear and concise, the authors employed several strategies. This included using straightforward, jargon-free language and peer review of the initial draft by the remainder of the author team to refine the wording. These steps were crucial in minimizing potential ambiguity in the criteria and enhancing inter-rater reliability (Dawson 2017). By focusing on clear communication and standardized evaluation guidelines, the rubric aims to achieve high levels of agreement among different raters, ensuring that the assessment of AI-generated feedback is both reliable and valid.

The final prompt

```
"model": "gpt-4",
"temperature": 0.5,
"messages": [
  {
   "role": "system",
   "content": "You are a teacher who teaches <User>, an accounting student. Your task is to provide feedback to the student on their answer to a question. Use plain English. Encourage self-assessment. For instance, encourage students to reflect on their answer, identify areas where they think they did well or struggled, and then compare their self-assessment with the feedback provided. Be specific and clear in your feedback, avoiding generic phrases like 'Good work!'. Be objective and use a growth mindset."
  },
  {
   "role": "user",
   "content": "Given the question: <Question>. Compare <User>'s answer to the teacher's answer. <User>'s answer: <Answer>. Teachers answer: <Memo>. The student has provided the following self-reflection: <Reflection>. The student would like feedback on the following specific aspects: <SpecificFeedback>."
  },
  {
   "role": "user",
   "content": "Draft a report to <User> providing feedback on their answer when compared to the teacher's answer. Start by acknowledging the effort put into the task and encouraging a growth mindset. Next, thoroughly analyse and carefully inspect the student's answer to identify specific strengths and weaknesses. Provide detailed and specific examples of these strengths and weaknesses in the student's solution. When identifying strengths, highlight the aspects of the answer that demonstrate a good theoretical knowledge and understanding of the material, as well as the ability to apply that knowledge in the given context. When identifying weaknesses, pinpoint specific areas where the answer could be improved, such as pointing out incorrect calculations, incomplete information, unclear explanations, or omissions of theory. Provide in-depth feedback by carefully unpacking the student's answer and addressing errors and omissions. Maintain a motivational and supportive tone throughout. Offer actionable suggestions to improve the identified weaknesses. Address the student's reflection and requested feedback, if any. Conclude by summarizing the improvement suggestions in a concise manner."
  },
  ]
}
```

Figure 2. The final prompt.

To quantify the evaluation, a 3-point Likert scale was adopted for its simplicity and ease of interpretation. This scale enabled the differentiation between feedback that does not meet, partially meets or fully meets the established criteria. The scale, by limiting options to three discrete categories, allows raters to concentrate on whether the feedback meaningfully supports learning outcomes, rather than fine distinctions that might not be as impactful on student improvement.

Results and discussion

The evaluation of the effectiveness of the prompt revealed a mean adherence to the effective feedback principles (Nicol and Macfarlane-Dick 2006) of 2.67 out of 3 (Tables 3 and 4). This shows that the LLM mostly provided clear, constructive feedback aligned with best practices.

Deeper analysis revealed some variability in the effectiveness of the AI-generated feedback across different principles, underscoring the complexity of creating a prompt that enables an LLM to consistently apply pedagogical best practices. For example, improvement is needed in consistently delivering high-quality information about students' learning and in facilitating the development of self-assessment (reflection) in learning. These principles consistently received the lowest ratings across all assessments (Table 3).

Table 1. Alignment between feedback principles and the prompt.

Feedback principle	Quote from the Prompt In *Bubble.io*, the notation "< >" is used to dynamically insert specific data into a text field or other component within an application. When you see "<User>", it signifies that the user's name will be automatically populated in that place.
Clarity of what good performance is (*i.e.* the goal)	Compare < User>'s answer to the teacher's answer. <User>'s answer: <Answer>. Teachers answer: <Memo>.
Facilitates the development of self-assessment (reflection) in learning	Encourage self-assessment. For instance, encourage students to reflect on their answer, identify areas where they think they did well or struggled, and then compare their self-assessment with the feedback provided. Address the student's reflection and requested feedback
Delivers high-quality information to students about their learning	Compare < User>'s answer to the teacher's answer. <User>'s answer: <Answer>. Teachers answer: <Memo>. Next, thoroughly analyse and carefully inspect the student's answer to identify specific strengths and weaknesses. Provide detailed and specific examples of these strengths and weaknesses in the student's solution. When identifying strengths, highlight the aspects of the answer that demonstrate a good theoretical knowledge and understanding of the material, as well as the ability to apply that knowledge in the given context. When identifying weaknesses, pinpoint specific areas where the answer could be improved, such as pointing out incorrect calculations, incomplete information, unclear explanations, or omissions of theory. Provide in-depth feedback by carefully unpacking the student's answer and addressing errors and omissions. Use plain English.
Encourages instructor and peer dialogue around learning	Maintain a motivational and supportive tone throughout.
Encourages positive motivational beliefs and self-esteem	Maintain a motivational and supportive tone throughout. Be specific and clear in your feedback, avoiding generic phrases like 'Good work!'. Be objective and use a growth mindset. Start by acknowledging the effort put into the task and encouraging a growth mindset.
Provides opportunities to close the gap between current and desired performance	Offer actionable suggestions to improve the identified weaknesses. Conclude by summarizing the improvement suggestions in a concise manner.
Provides information to instructors that can be used to help shape teaching.	*Not directly applicable to the prompt. The web application provides information to instructors via a separate action.*

Feedback principle 1: clarity of what good performance is (i.e. the goal)

The evaluation of feedback principle 1 indicates that while the LLM frequently leveraged the suggested solutions to guide its feedback ($n = 59$), there were some instances where the responses omitted explicit comparisons ($n = 12$). The absence of specific comparative references in some feedback instances might hinder students' ability to self-evaluate accurately. Future iterations of the underlying LLM such as GPT-5 may bolster this aspect. In the interim, where possible, to ensure that students are aware of what good performance is, instructors should provide students with clear criteria for what constitutes good performance. They might also consider supplementing AI-generated feedback with examples of high-quality answers or offering brief workshops on effective self-assessment techniques. This approach could help mitigate the current limitations of AI-generated feedback, enhancing students' learning experiences by fostering a deeper understanding and application of course concepts.

Feedback principle 2: facilitates the development of self-assessment (reflection) in learning

The web application invites students to reflect on their answers and specify areas where they seek feedback. Although the study design intentionally selected the 'none' option for these questions to focus on evaluating direct feedback effectiveness, the AI's proactive stance, evident in further encouraging self-evaluation, was notable. In the majority of instances, the LLM initiated additional prompts for self-reflection (some responses had more than one prompt), highlighting the importance of self-assessment in the learning journey. For instance, feedback included

Table 2. Principles of effective feedback.

	Panel A: Feedback Principle	Criteria	Rating Scale (1–3)
1	Clarity of what good performance is (i.e. the goal)	• The student's answer was compared to the instructor's model answer. • The feedback clearly communicates the expectations and standards of the task based on the instructor's model answer.	1 = Principle not adhered to 2 = Principle partially adhered to 3 = Principle adhered to
2	Facilitates the development of self-assessment (reflec-tion) in learning	• The student was asked what they want feedback on. • The student was encouraged to reflect and self-assess by asking them what *they* think the strengths and weaknesses are of their answer, and where they struggled. • The feedback promotes active engagement in the feedback process with the student.	1 = Principle not adhered to 2 = Principle partially adhered to 3 = Principle adhered to
3	Delivers high-quality information to students about their learning	• The feedback is technically accurate. • The feedback relates to a specific standard of performance. • The feedback uses plain English. Plain English refers to the use of straightforward and clear language that avoids jargon, complex sentence structures, and technical terms that are not easily understood by a general audience. Plain English emphasizes directness, simplicity, and clarity, making essential concepts and instructions accessible to the student. • The feedback highlights specific strengths and weaknesses in the students' work. • The feedback offers actionable suggestions for improvement. • - The feedback helps the student understand how they can apply what they have learned or improve in the next assessment.	1 = Principle not adhered to 2 = Principle partially adhered to 3 = Principle adhered to
4	Encourages instructor and peer dialogue around learning	• The feedback fosters a dialogue between the AI and the student. • The feedback creates a positive environment for the dialogue.	1 = Principle not adhered to 2 = Principle partially adhered to 3 = Principle adhered to
5	Encourages positive motivational beliefs and self-esteem	• The feedback maintains a motivational tone and encourages a growth mindset by focusing on the students' progress and how they can improve. • The feedback supports the development of positive self-esteem and motivation in the student.	1 = Principle not adhered to 2 = Principle partially adhered to 3 = Principle adhered to
6	Provides opportunities to close the gap between current and desired performance	• The feedback offers actionable suggestions for improvement. • The feedback enables the student to understand how they can apply what they have learned or improve in the next assessment.	1 = Principle not adhered to 2 = Principle partially adhered to 3 = Principle adhered to
7	Provides information to instructors that can be used to help shape teaching	• The feedback generates information that can help instructors understand the students' progress and adjust their teaching methods accordingly.	1 = Principle not adhered to 2 = Principle partially adhered to 3 = Principle adhered to

Table 3. Mean scores of the presence of effective feedback principles per assessment.

	Feedback principle (mean)							
Assessment	1	2	3	4	5	6	7	Overall mean per assessment
1	2.64	2.36	2.00	3.00	3.00	2.57	3.00	2.65
2	2.92	2.46	2.00	3.00	3.00	2.62	3.00	2.71
3	2.94	2.28	2.00	3.00	3.00	2.50	3.00	2.67
4	2.77	2.31	2.00	3.00	3.00	2.85	3.00	2.70
5	2.85	2.15	2.00	3.00	3.00	2.23	3.00	2.60
Overall mean per feedback principle	2.83	2.31	2.00	3.00	3.00	2.55	3.00	2.67

1 – Clarity of what good performance is (i.e. the goal); 2 – Facilitates the development of self-assessment (reflection) in learning; 3 – Delivers high-quality information to students about their learning; 4 – Encourages instructor and peer dialogue around learning; 5 – Encourages positive motivational beliefs and self-esteem; 6 – Provides opportunities to close the gap between current and desired performance; 7 – Provides information to instructors that can be used to help shape teaching.

encouragements like '*Remember, the goal is to deepen your understanding of the intricacies of accounting principles and their application, and self-reflection plays a key role in doing that*' and advice to '*write down your thoughts after writing an answer. What parts were easy? Where did you struggle? This will help me provide you with more targeted feedback*'. These prompts demonstrate a sophisticated approach to embedding reflective learning processes within feedback mechanisms. This aspect the feedback system aligns with educational best practices, emphasizing the value of reflection in enhancing student understanding and performance (Nicol and Macfarlane-Dick 2006).

Feedback principle 3: delivers high-quality information to students about their learning

Most of the feedback (59 out of 71 responses) effectively drew on suggested solutions for constructive comparison. The AI-generated feedback's detailed structure – highlighting strengths before suggesting improvements – mirrors best practices in educational feedback. For example, '*your discussion on faithful representation was strong. You've accurately identified that the capitalization of the expenses led to incomplete information, lack of neutrality, and a presentation that was not free from error',* followed by comments such as '*However, when comparing your answer to the detailed feedback from the teacher, there are a few areas that could be improved or clarified*'.

However, despite the LLM's capability to provide nuanced evaluations, discrepancies in the depth of analysis, occasional failure to identify key errors and instances of over-praise were identified. To address these variations in feedback, instructors can adopt several strategies such as reviewing critical assignments, hosting workshops on self-assessment, guiding students through reflective questioning, fostering feedback literacy, employing a hybrid feedback model and facilitating peer discussions.

LLM-generated feedback should ideally be reviewed and, if necessary, edited by an instructor before being shared with students to ensure its accuracy, relevance and appropriateness (Nysom 2023). Given that this might not always be possible in large classes, it's important to make students aware of the potential biases and errors in the feedback (Meyer et al. 2024), and to encourage them to assess and confirm the feedback from LLMs with reliable sources (Lo 2023).

As another consideration in the provision of high-quality feedback, the AI-generated feedback was assessed for readability using the Flesch Reading Ease score in *Microsoft Word*. Flesch's reading ease standards suggest that scores above 50 are generally easier to understand, and those of 30 or below can be challenging, aligning with college-level difficulty (Courtis and Hassan 2002). While lower scores on this scale indicate texts that are more difficult to comprehend (Courtis 2004), university materials are expected to exhibit a certain level of sophistication. The AI-generated feedback average score of 31.11(Table 4), suggests it strikes a balance between complexity and comprehensibility, suitable for university standards.

Feedback principle 4: encouragement of instructor and peer dialogue around learning

The web application utilizes an instant messaging interface to promote interactive exchanges between students and the AI, simulating a conversational environment conducive to deeper engagement. After receiving initial feedback, students are encouraged to ask questions or express concerns, fostering a learning dialogue.

In 15 instances, the LLM proactively invited further interaction, underscoring the system's potential to stimulate reflective thinking and peer discussions. Messages like, '*Please let me know if you want to discuss specific parts of this feedback*' or '*Please don't hesitate to reach out if you have any further doubts or queries*' exemplify the application's design to encourage a dialogic learning process.

To augment this AI-driven dialogue, instructors play a crucial role in facilitating deeper conversations. They can leverage AI-generated feedback as a springboard for group discussions or one-on-one sessions, focusing on areas where students commonly seek clarification. This not only enhances the feedback loop but also provides opportunities for peer learning and instructor-led guidance, vital for

Table 4. Flesch pattern of reading ease scores (Flesch 1948).

Flesch Reading Ease score	Description of reading level	*n*	%
0–30	Very difficult	24	34%
30–50	Difficult	47	66%
Total		71	100%
Average		31.61	
Median		34.10	

addressing nuances the LLM might miss. Implementing regular review sessions based on AI-generated feedback themes can further solidify understanding and application of course materials.

Instructors monitoring these interactions gain valuable insights into common student challenges, informing teaching strategies and potential AI prompt adjustments. By strategically integrating AI feedback into the broader educational dialogue, educators can significantly enrich the learning experience, promoting a culture of continuous improvement and reflection.

Feedback principle 5: encourages positive motivational beliefs and self-esteem

The LLM consistently generated feedback responses that supported a positive learning environment by including affirming phrases to motivate students. Phrases such as '*Keep up the great work and keep learning*!' and '*You're on the right track and I encourage you to continue to work hard*' were common, serving not only to reinforce effort but also to cultivate a growth mindset essential for academic progress. This constructive feedback began by acknowledging what students did well, providing immediate positive reinforcement, and then offered specific suggestions for improvement. This specificity avoided generic criticism, facilitating a clear understanding of both strengths and areas needing attention, and outlined actionable steps for students to enhance their work. By encouraging students to reflect on their growth, the feedback from the LLM can play a pivotal role in fostering continuous engagement and self-assessment, aligning with effective pedagogical practices for nurturing motivation and self-esteem.

Feedback principle 6: provides opportunities to close the gap between current and desired performance

The analysis revealed that a significant proportion of the feedback responses ($n = 39$) offered practical suggestions, guiding students toward closing the performance gap to reach their goals. Other responses ($n = 24$), while offering suggestions and guiding students toward closing the performance gap to reach their goals, provided recommendations that were either incomplete or were misaligned with the students' answers, being either incorrect or irrelevant. Certain of these feedback responses implied suggestions through the critique of weaknesses, without directly stating them, or were very specific to the question at hand. This implicit and/or specific advice, while tailored to the specific query, may challenge students' ability to generalize the underlying principles to different contexts, a skill associated with higher-order thinking (Lewis and Smith 1993). A small number of feedback responses ($n = 8$) were identified as vague, potentially leaving students uncertain about how to implement the advice provided.

Feedback principle 7: provides information to instructors that can be used to help shape teaching

As the tool is self-developed, the instructors have the capability to review feedback responses provided to students. This feature facilitates the use of feedback as a diagnostic instrument, allowing instructors to gauge student progress and identify areas requiring additional support or clarification. The accessibility of feedback transcripts empowered the instructors to adapt their teaching strategies based on real-time insights into student understanding and performance.

The observed variations in feedback quality (Table 3) point to the AI's fluctuating interpretive capacities, which can depend on the complexity and specificity of the student inputs. Such

fluctuations underscore the importance of human oversight in the feedback loop, ensuring that AI-generated advice aligns with pedagogical objectives.

The potential of LLMs such as ChatGPT to enhance educational processes must also be weighed against the concerns of broader automation and the increased risk of its misuse by students. To mitigate these concerns, it is imperative to preserve the human elements of education, ensuring that LLMs such as Chat GPT serve as a complementary tool rather than a replacement for human judgment and interaction (Baidoo-Anu and Owusu Ansah 2023; Grassini 2023; Kasneci et al. 2023). The possibility of an error rate in AI-generated feedback leading to negative consequences for students is an ethical concern. However, just as human instructors are subject to rigorous quality control to ensure the accuracy, relevance, consistency and effectiveness of their feedback, so too should AI-generated feedback be carefully reviewed and monitored (Jacobsen and Weber 2023). Maintaining instructor oversight in the feedback process is crucial to ensure the reliability of both AI-generated and human-provided feedback, safeguarding student well-being and academic integrity.

Limitations and future research

Despite these promising results, the study has several limitations. The evaluation focused solely on second-year accounting students, which may limit the generalizability of the findings to other disciplines or contexts. The study did not engage further with the feedback provided by the LLM. Future research could investigate the impact of direct engagement with reflective prompts on student learning outcomes and well-being, offering deeper insights into the role of AI in supporting metacognitive skills development. The impact of the feedback given by the LLM on the students learning experience or performance on subsequent tasks was also not investigated. Future studies could explore students' perceptions of the feedback generated by the LLM, or whether the regular use of AI generated feedback systems can impact students' learning. Future work could also focus on identifying the conditions under which an LLM performs best and where human intervention is most critical. These findings could then inform the development of more sophisticated LLMs tailored to educational feedback. Future work should also focus on developing advanced error-checking mechanisms within LLM systems and investigating the broader ethical implications of LLMs in educational settings (Lo 2023).

Conclusion

This paper highlights the significant challenges inherent in traditional feedback methods within large educational contexts, particularly the intensive time and resource requirements needed to provide personalized, constructive feedback. To address these challenges, this paper explored the integration of AI, focusing on the capabilities of OpenAI's GPT-4 LLM. By developing a prompt grounded in the theoretical framework of effective feedback proposed by Nicol and Macfarlane-Dick (2006), this paper aimed to effectively streamline the feedback process in large educational contexts while addressing the limitations and ethical considerations related to AI inclusion.

The integration of LLMs into educational feedback processes presents both opportunities and challenges. The evaluation of AI-generated feedback showed a promising alignment with the principles of effective feedback, suggesting that LLMs can significantly enhance the scalability and consistency of feedback, particularly in large classes where personalized feedback is traditionally difficult to deliver. However, the study also uncovered instances where the feedback generated by the LLM diverged from these principles, with missed opportunities to identify misconceptions or provide comprehensive, actionable advice. This underscores the necessity of balancing sophistication with accessibility to ensure feedback remains both intellectually stimulating and comprehensible.

LLMs should, at present, be viewed as a tool that complements human instructors rather than replacing them (Grassini 2023). The ethical rationale for incorporating AI in education lies in its potential to not only enhance instructors' capabilities to manage large classes, provide personalized feedback at scale, and identify patterns that might escape human observation, but also the potential to develop students' ability to question and critically analyse content, a key twenty first century skill (Farrelly and Baker 2023).

By maintaining a balance between AI integration and human oversight, the feedback process can benefit from the efficiency of LLMs while safeguarding its integrity and human-centric values. Ultimately, AI should reinforce the human-centered nature of education, not undermine it.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Julia Venter is conducting a Ph.D. study at the Department of Accounting at the University of Pretoria (UP), South Africa on the topic of "incorporating social constructivism in introductory accounting".

Stephen A. Coetzee, PhD, is a professor at the University of Pretoria, South Africa. His research interest is competency-based accounting education.

Astrid Schmulian, PhD, is an associate professor at the University of Pretoria, South Africa. Her research interest is competency-based accounting education.

ORCID

Stephen A. Coetzee http://orcid.org/0000-0003-3092-0257
Astrid Schmulian http://orcid.org/0000-0003-4946-3076
Julia Venter http://orcid.org/0000-0001-8666-231X

References

Alshater, Muneer. 2022. "Exploring the Role of Artificial Intelligence in Enhancing Academic Performance: A Case Study of ChatGPT." *SSRN Electronic Journal* 2022, 4312358. doi:10.2139/ssrn.4312358.

Baidoo-Anu, David, and Leticia Owusu Ansah. 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. doi:10.61969/jai.1337500.

Balloo, Kieran, Carol Evans, Annie Hughes, Xiaotong Zhu, and Naomi Winstone. 2018. "Transparency Isn't Spoon-Feeding: How a Transformative Approach to the Use of Explicit Assessment Criteria Can Support Student Self-Regulation." *Frontiers in Education* 3: 69. doi:10.3389/feduc.2018.00069.

Bauer, Elisabeth, Martin Greisel, Ilia Kuznetsov, Markus Berndt, Ingo Kollar, Markus Dresel, Martin R. Fischer, and Frank Fischer. 2023. "Using Natural Language Processing to Support Peer-Feedback in the Age of Artificial Intelligence: A Cross-Disciplinary Framework and a Research Agenda." *British Journal of Educational Technology* 54 (5): 1222–1245. doi:10.1111/bjet.13336.

Beatman, Andy. 2023. "Prompts are key in 2023: Twenty-five tips to help you unlock the potential of generative AI." *Microsoft*. https://azure.microsoft.com/en-us/blog/prompts-are-key-in-2023-twenty-five-tips-to-help-you-unlock-the-potential-of-generative-ai/.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021, March. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bommarito, Jillian, Michael Bommarito, Daniel M. Katz, and Jessica Katz. 2023. "GPT as Knowledge Worker: A Zero-Shot Evaluation of (AI) CPA Capabilities." arXiv. *preprint arXiv:2301.04408*. doi:10.48550/arXiv.2301.04408.

Borji, Ali. 2023. "A Categorical Archive of ChatGPT Failures." arXiv. *2302.03494v8*. doi:10.48550/arXiv.2302.03494.

Boud, David. 2000. "Sustainable Assessment: Rethinking Assessment for the Learning Society." *Studies in Continuing Education* 22 (2): 151–167. doi:10.1080/713695728.

Boud, David, Rosemary Keogh, and David Walker. 2013. *Reflection: Turning Experience into Learning*. London: Routledge.

Boud, David, and Elizabeth Molloy. 2013. "Rethinking Models of Feedback for Learning: The Challenge of Design." *Assessment & Evaluation in Higher Education* 38 (6): 698–712. doi:10.1080/02602938.2012.691462.

Butler, Deborah L., and Philip H. Winne. 1995. "Feedback and Self-Regulated Learning: A Theoretical Synthesis." *Review of Educational Research* 65 (3): 245–281. doi:10.3102/00346543065003245.

Carless, David, and David Boud. 2018. "The Development of Student Feedback Literacy: Enabling Uptake of Feedback." *Assessment & Evaluation in Higher Education* 43 (8): 1315–1325. doi:10.1080/02602938.2018.1463354.

Carless, David, and Naomi Winstone. 2023. "Teacher Feedback Literacy and Its Interplay with Student Feedback Literacy." *Teaching in Higher Education* 28 (1): 150–163. doi:10.1080/13562517.2020.1782372.

Cavalcanti, Anderson, Rafael Ferreira Mello, Vitor Rolim, Máverick André, Fred Freitas, and Dragan Gaševic. 2019. "An Analysis of the Use of Good Feedback Practices in Online Learning Courses." 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), 15–18 July 2019.

Cavalcanti, Anderson, Arthur Diego, Rafael Ferreira Mello, Katerina Mangaroska, André Nascimento, Fred Freitas, and Dragan Gašević. 2020. "How Good is my Feedback? A Content Analysis of Written Feedback." Proceedings of the Tenth International Conference on Learning Analytics & Knowledge.

Chomsky, Noam, Ian Roberts, and Jeffrey Watumull. 2023. "The False Promise of ChatGPT." *The New York Times*, 8 March 2023, 2023. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

Courtis, John K. 2004. "Corporate Report Obfuscation: Artefact or Phenomenon?" *The British Accounting Review* 36 (3): 291–312. doi:10.1016/j.bar.2004.03.005.

Courtis, John K., and Salleh Hassan. 2002. "Reading Ease of Bilingual Annual Reports." *Journal of Business Communication* 39 (4): 394–413. doi:10.1177/002194360203900401.

Dai, Wei, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. "Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT." 2023 IEEE International Conference on Advanced Learning Technologies (ICALT). doi:10.1109/ICALT58122.2023.00100.

Dawson, Phillip. 2017. "Assessment Rubrics: Towards Clearer and More Replicable Design, Research and Practice." *Assessment & Evaluation in Higher Education* 42 (3): 347–360. doi:10.1080/02602938.2015.1111294.

Deeva, Galina, Daria Bogdanova, Estefanía Serral, Monique Snoeck, and Jochen De Weerdt. 2021. "A Review of Automated Feedback Systems for Learners: Classification Framework, Challenges and Opportunities." *Computers & Education* 162: 104094. doi:10.1016/j.compedu.2020.104094.

Demszky, Dorottya, Jing Liu, Heather C. Hill, Dan Jurafsky, and Chris Piech. 2023. "Can Automated Feedback Improve Teachers' Uptake of Student Ideas? evidence from a Randomized Controlled Trial in a Large-Scale Online Course." *Educational Evaluation and Policy Analysis* 46 (3): 483–505. doi:10.3102/01623737231169270.

Dweck, Carol S. 2000. *Self-Theories: Their Role in Motivation, Personality, and Development*. Philadelphia: Psychology Press.

Evans, Carol. 2013. "Making Sense of Assessment Feedback in Higher Education." *Review of Educational Research* 83 (1): 70–120. doi:10.3102/00346543124743.

Farrelly, Tom, and Nick Baker. 2023. "Generative Artificial Intelligence: Implications and Considerations for Higher Education Practice." *Education Sciences* 13 (11): 1109. doi:10.3390/educsci13111109.

Fisher, Douglas, and Nancy Frey. 2009. "Feed up, Back, Forward." *Educational Leadership* 67 (3): 20–25. https://fisher-and-frey.s3.amazonaws.com/documents/feed_forward.pdf.

Flesch, R. 1948. "A New Readability Yardstick." *Journal of Applied Psychology* 32 (3): 221–233. doi:10.1037/h0057532.

Gardner, John, Michael O'Leary, and Li Yuan. 2021. "Artificial Intelligence in Educational Assessment: 'Breakthrough? Or Buncombe and Ballyhoo?" *Journal of Computer Assisted Learning* 37 (5): 1207–1216. doi:10.1111/jcal.12577.

Gold, Raymond L. 1958. "Roles in Sociological Field Observations." *Social Forces* 36 (3): 217–223. http://www.jstor.org/stable/2573808. doi:10.2307/2573808.

Grassini, Simone. 2023. "Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings." *Education Sciences* 13 (7): 692. doi:10.3390/educsci13070692.

Hattie, John. 2008. *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement*. Abingdon: Routledge.

Hattie, John, and Helen Timperley. 2007. "The Power of Feedback." *Review of Educational Research* 77 (1): 81–112. doi:10.3102/003465430298487.

Henderson, Michael, Michael Phillips, Tracii Ryan, David Boud, Phillip Dawson, Elizabeth Molloy, and Paige Mahoney. 2019. "Conditions That Enable Effective Feedback." *Higher Education Research & Development* 38 (7): 1401–1416. doi:10.1080/07294360.2019.1657807.

Higgins, Richard, Peter Hartley, and Alan Skelton. 2001. "Getting the Message Across: The Problem of Communicating Assessment Feedback." *Teaching in Higher Education* 6 (2): 269–274. doi:10.1080/13562510120045230.

Hooda, Monika, Chhavi Rana, Omdev Dahiya, Ali Rizwan, and Md Shamim Hossain. 2022. "Artificial Intelligence for Assessment and Feedback to Enhance Student Success in Higher Education." *Mathematical Problems in Engineering* 2022: 1–19. doi:10.1155/2022/5215722.

Jacobsen, Lucas Jasper, and Kira Elena Weber. 2023. "The Promises and Pitfalls of ChatGPT as a Feedback Provider in Higher Education: An Exploratory Study of Prompt Engineering and the Quality of AI-Driven Feedback." OSF Preprints. doi:10.31219/osf.io/cr257.

Kasneci, Enkelejda, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, et al. 2023. "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education." *Learning and Individual Differences* 103: 102274. doi:10.1016/j.lindif.2023.102274.

Kirk, James R., Robert E. Wray, Peter Lindes, and John E. Laird. 2022. "Improving language model prompting in support of semi-autonomous task learning." *arXiv preprint arXiv:2209.07636*. doi:10.48550/arXiv.2209.07636.

Kung, Tiffany H., Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, et al. 2023. "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models." *PLoS Digital Health* 2 (2): e0000198. doi:10.1371/journal.pdig.0000198.

Laurillard, Diana. 2013. *Rethinking University Teaching: A Conversational Framework for the Effective Use of Learning Technologies*. Abingdon: Routledge.

Lewis, Arthur, and David Smith. 1993. "Defining Higher Order Thinking." *Theory Into Practice* 32 (3): 131–137. doi:10.1080/00405849309543588.

Lo, Chung Kwan. 2023. "What is the Impact of ChatGPT on Education? A Rapid Review of the Literature." *Education Sciences* 13 (4): 410. doi:10.3390/educsci13040410.

Marcus, G., and E. Davis. 2020. "GPT-3, Bloviator: OpenAI's Language Generator Has No Idea What It's Talking About." MIT Technology Review. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/.

Meyer, Jennifer, Thorben Jansen, Ronja Schiller, Lucas W. Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. "Using LLMs to Bring Evidence-Based Feedback into the Classroom: AI-Generated Feedback Increases Secondary Students' Text Revision, Motivation, and Positive Emotions." *Computers and Education: Artificial Intelligence* 6: 100199. doi:10.1016/j.caeai.2023.100199.

Nicol, David J., and Debra Macfarlane-Dick. 2006. "Formative Assessment and Self-Regulated Learning: A Model and Seven Principles of Good Feedback Practice." *Studies in Higher Education* 31 (2): 199–218. doi:10.1080/03075070600572090.

Nunes, Andreia, Carolina Cordeiro, Teresa Limpo, and São Luís Castro. 2022. "Effectiveness of Automated Writing Evaluation Systems in School Settings: A Systematic Review of Studies from 2000 to 2020." *Journal of Computer Assisted Learning* 38 (2): 599–620. doi:10.1111/jcal.12635.

Nysom, Lars. 2023. "AI Generated Feedback for Students' Assignment Submissions." [Master's thesis, Aalborg University]. Aalborg University Project Library. https://projekter.aau.dk/projekter/files/547261577/Lars_Nysom_Master_Project.pdf

OpenAI. n.d. "API Documentation." Accessed 28 April 2023. https://platform.openai.com/overview.

Orsmond, Paul, Stephen Merry, and Kevin Reiling. 2002. "The Use of Exemplars and Formative Feedback When Using Student Derived Marking Criteria in Peer and Self-Assessment." *Assessment & Evaluation in Higher Education* 27 (4): 309–323. doi:10.1080/0260293022000001337.

Pardo, Abelardo, Jelena Jovanovic, Shane Dawson, Dragan Gašević, and Negin Mirriahi. 2019. "Using Learning Analytics to Scale the Provision of Personalised Feedback." *British Journal of Educational Technology* 50 (1): 128–138. doi:10.1111/bjet.12592.

Parikh, Amish, Kylen McReelis, and Brian Hodges. 2001. "Student Feedback in Problem Based Learning: A Survey of 103 Final Year Students across Five Ontario Medical Schools." *Medical Education* 35 (7): 632–636. doi:10.1046/j.1365-2923.2001.00994.x.

Pegoraro, Alessandro, Kavita Kumari, Hossein Fereidooni, and Ahmad-Reza Sadeghi. 2023. "To ChatGPT, or not to ChatGPT: That is the question!" *arXiv preprint arXiv:2304.01487*. doi:10.48550/arXiv.2304.01487.

Pinto, Gustavo, Isadora Cardoso-Pereira, Danilo Monteiro, Danilo Lucena, Alberto Souza, and Kiev Gama. 2023. "Large Language Models for Education: Grading Open-Ended Questions Using Chatgpt." Proceedings of the XXXVII Brazilian Symposium on Software Engineering. doi:10.1145/3613372.3614197.

Ramesh, Dadi, and Suresh Kumar Sanampudi. 2022. "An Automated Essay Scoring Systems: A Systematic Literature Review." *Artificial Intelligence Review* 55 (3): 2495–2527. doi:10.1007/s10462-021-10068-2.

Ray, Partha Pratim. 2023. "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope." *Internet of Things and Cyber-Physical Systems* 3: 121–154. doi:10.1016/j.iotcps.2023.04.003.

Rupp, André A., Jodi M. Casabianca, Maleika Krüger, Stefan Keller, and Olaf Köller. 2019. "Automated Essay Scoring at Scale: A Case Study in Switzerland and Germany." *ETS Research Report Series* 2019 (1): 1–23. doi:10.1002/ets2.12249.

Sadler, D. Royce. 1989. "Formative Assessment and the Design of Instructional Systems." *Instructional Science* 18 (2): 119–144. doi:10.1007/BF00117714.

Schramowski, Patrick, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. 2022. "Large Pre-Trained Language Models Contain Human-like Biases of What is Right and Wrong to Do." *Nature Machine Intelligence* 4 (3): 258–268. doi:10.1038/s42256-022-00458-8.

Sok, Sarin, and Kimkong Heng. 2023. "ChatGPT for Education and Research: A Review of Benefits and Risks." *Cambodian Journal of Educational Research* 3 (1): 110–121. https://www.researchgate.net/profile/Cambodian-Journal-Of-Educational-Research/publication/373170005_Cambodian_Journal_of_Educational_Research_Volume_3_Number_1/links/64de1f111351f5785b707247/Cambodian-Journal-of-Educational-Research-Volume-3-Number-1.pdf#page=129. doi:10.62037/cjer.2023.03.01.06.

Tekian, Ara, Christopher J. Watling, Trudie E. Roberts, Yvonne Steinert, and John Norcini. 2017. "Qualitative and Quantitative Feedback in the Context of Competency-Based Education." *Medical Teacher* 39 (12): 1245–1249. doi :10.1080/0142159X.2017.1372564.

Terwiesch, Christian. 2023. "Would Chat GPT3 get a Wharton MBA? A prediction based on its performance in the operations management course." *Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania* 45. https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Would-ChatGPT-get-a-Wharton-MBA.pdf.

van de Pol, Janneke, Monique Volman, and Jos Beishuizen. 2010. "Scaffolding in Teacher–Student Interaction: A Decade of Research." *Educational Psychology Review* 22 (3): 271–296. doi:10.1007/s10648-010-9127-6.

White, Jules, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. "A prompt pattern catalog to enhance prompt engineering with chatgpt." "arXiv." *preprint arXiv:2302.11382*. doi:10.48550/arXiv.2302.11382.

Winstone, Naomi E., Robert A. Nash, James Rowntree, and Michael Parker. 2017. "It'd Be Useful, but I Wouldn't Use It': barriers to University Students' Feedback Seeking and Recipience." *Studies in Higher Education* 42 (11): 2026–2041. doi:10.1080/03075079.2015.1130032.

Yang, Min, and David Carless. 2013. "The Feedback Triangle and the Enhancement of Dialogic Feedback Processes." *Teaching in Higher Education* 18 (3): 285–297. doi:10.1080/13562517.2012.719154.

Yorke, Mantz. 2003. "Formative Assessment in Higher Education: Moves towards Theory and the Enhancement of Pedagogic Practice." *Higher Education* 45 (4): 477–501. doi:10.1023/A:1023967026413.

Zhang, Haoran, Ahmed Magooda, Diane Litman, Richard Correnti, Elaine Wang, L. C. Matsmura, Emily Howe, and Rafael Quintana. 2019. "eRevise: Using Natural Language Processing to Provide Formative Feedback on Text Evidence Usage in Student Writing." Proceedings of the AAAI Conference on Artificial Intelligence. doi:10.1609/aaai.v33i01.33019619.