# Exploring the use of AI-augmented feedback in written assignments

Jiayi Wang[ab]* and C. I. Galliou[ac]

[a]UCL Institute of Education, University College London, London, UK

[b]ORCID: 0009-0003-9547-5681

[c]ORCID: 0009-0009-2489-1235

*jiayi.w@ucl.ac.uk

# Exploring the use of AI-augmented feedback in written assignments


## Abstract

Generative Artificial Intelligence (GenAI) has rapidly emerged as a transformative force in education. While its potential benefits are vast, substantial gaps remain in understanding its role in transforming feedback practices. This study explored AI-augmented feedback in master's-level written assignments, analysing eight assignments with lecturer-provided feedback across two phases. Using Microsoft Copilot, the study examined differences between human-generated and AI-augmented feedback, highlighted strengths and limitations, and proposed strategies to optimise AI feedback generation. The findings revealed that AI-augmented feedback effectively addressed surface-level issues, such as grammar and formatting, and improved clarity and consistency. Motivational language and structured improvements aligned with effective feedback practices, enhancing engagement. However, the absence of targeted prompts often resulted in generic feedback, lacking depth and actionable insights. Limitations included reliance on templated responses, occasional inappropriate word choices, and a limited capacity to critique complex ideas. The study highlights the importance of refining prompt design to maximise the potential of AI in generating meaningful feedback. Ongoing research could focus on exploring larger, more diverse samples and updated AI models to evaluate long-term impacts and develop standardised practices for AI-augmented feedback.


**Keywords:** artificial intelligence in education; ai-augmented feedback; Microsoft Copilot; written assignments

## Introduction

The emergence of generative Artificial Intelligence (GenAI) in recent years, as one of the most significant innovations in human history, has sparked widespread discussion across various sectors, including education (Miao et al., 2021; Selwyn, 2022). AI-powered tools, driven by rapid technological advancement, have been positioned at the forefront of contemporary educational discourse, shaping research initiatives, informing policy development, and transforming teaching and learning practices (Eager & Brunton, 2023; Giannakos et al., 2024). Given the long-standing emphasis on the importance of effective feedback in promoting student learning and academic performance (Brookhart, 2017; Dawson et al., 2018; Hattie &

Timperley, 2007), a growing interest has emerged in integrating AI tools to augment feedback processes. AI-augmented feedback involves feedback initially generated by a human expert and subsequently processed by an AI tool to improve its clarity and motivational quality (Schultze et al., 2024). However, several constraints and challenges associated with providing reliable AI-augmented feedback (Schultze et al., 2024) and AI-generated responses (Groves et al., 2024; Misanchuk & Hyzyk, 2024; Obaidoon & Wei, 2024) have been widely recognised and discussed.

This study was initiated as part of the AI & Assessment Challenge organised by UCL Institute of Education (IOE), where participants were briefed on several pressing AI-related issues and provided with resources to explore how AI tools could transform academic practices. Inspired by Schultze et al.'s (2024) suggestion that AI-augmented feedback could address challenges related to feedback quality in higher education, this study focused on its application in academic written assignments at the master's level. The following research questions were posed:

1. What are the differences between the original feedback provided by lecturers and AI-augmented feedback for academic written assignments?
2. What are the strengths and limitations of using AI-augmented feedback in academic written assignments?
3. What strategies could be employed to further promote the use of AI-augmented feedback?

In this study, Microsoft Copilot was used to generate AI-augmented feedback, as recommended by the Challenge guidelines and due to its relative lack of research in educational contexts compared to other GenAI tools employing large language models (LLMs), such as ChatGPT. Microsoft Copilot is also designed to integrate smoothly with widely used Microsoft applications, including Word, Excel, PowerPoint, Outlook, and Teams (Spataro, 2023), making it more accessible for educators to generate AI-augmented feedback directly within these platforms, thereby streamlining and simplifying their workflows.

By examining the differences between human-generated and AI-augmented feedback, this study contributes to the growing body of literature on AI in education, highlighting the strengths and limitations of integrating AI-augmented feedback to enhance the quality of feedback provided to students in academic contexts. The findings also provide practical insights into optimising prompts for generative AI tools to improve the quality and efficacy of AI-augmented feedback. Furthermore, the study lays a foundation for future research with larger sample sizes to further explore AI's broader role, applications, and long-term impacts in the field of education.

## Literature review

Feedback has long been a critical component of education, serving as a key mechanism for guiding learning and fostering improvement (Hattie & Timperley, 2007). Over time, changes in pedagogical theories, social expectations, and technological advancements have transformed perceptions of feedback. Historically, feedback was primarily regarded as a corrective tool (Nicol, 2010). However, contemporary perspectives frame feedback as a relational and dialogic process (Carless, 2015), emphasising its role in fostering collaboration, promoting learner growth, and enhancing student agency through formative practices (Nicol & Macfarlane-Dick, 2006). This process is achieved when feedback addresses essential questions about learning goals, current progress, and the actions required for improvement, thereby guiding learners toward achieving their objectives (Hattie & Timperley, 2007; Hattie & Yates, 2014).

In higher education, feedback is often categorised as formative or summative. Formative feedback is developmental, ongoing, and provides students with guidance to refine their skills throughout their learning journey (Carless, 2015). Summative feedback, by contrast, focuses on the final evaluation of overall performance, often using rubrics to ensure transparency and consistency in assessments (Carless, 2015).

### *Effective feedback*

Effective feedback is crucial for motivating students, fostering self-assessment, and improving their learning strategies (Hattie & Timperley, 2007). In higher education, although students recognise feedback as essential to their learning journey, they often struggle to engage with feedback that is unclear or generic (Glazzard & Stones, 2019), reducing its impact on their academic development (Lynam & Cachia, 2018). For feedback to be perceived as meaningful, it must be timely, personal, and detailed, offering targeted guidance for improvement (Beaumont et al., 2008; Glazzard & Stones, 2019; Hattie & Timperley, 2007), enabling students to act effectively on it and make improvements in subsequent assessments (Lynam & Cachia, 2018). Similarly, Dawson et al. (2018) highlight that students value feedback that is clear, personalised, and includes actionable suggestions for improvement. Shute (2008) further emphasises the importance of tailoring feedback to align with learners' characteristics, goals, and contexts. Providing feedback that references clear criteria and standards can help students understand expected performance levels, enhancing learning outcomes and fostering self-regulation (Nicol & Macfarlane-Dick, 2006). However, staff and students often refer to other elements as key to effective feedback, such as usability and specificity, over explicit links to standards (Dawson et al., 2018).

### *The use of AI in education*

Artificial Intelligence (AI) is a field dedicated to developing systems capable of replicating human cognitive processes, such as reasoning, learning, and adapting to new situations (Giannakos et al., 2024). In the context of education, AI systems are often perceived as tools for addressing complex learning challenges or performing tasks through capabilities traditionally associated with human intelligence, such as adapting to individual needs and providing interactive guidance (Chassignol et al., 2018). Technologies within the AI field, including machine learning, natural language processing (NLP), neural networks, and data mining (Baker & Smith, 2019, as cited in Guan et al., 2020), can contribute to advanced problem-solving capabilities (Guan et al., 2020).

### *Microsoft Copilot*

Microsoft Copilot is considered effective in addressing linguistic and structural challenges in writing due to its advanced natural language processing capabilities (Groves et al., 2024). Esfandiari and Allaf-Akbary (2024) further highlight its utility in identifying interactional metadiscourse issues for English as a Foreign Language (EFL) learners, thereby improving coherence and argumentation clarity in academic texts. However, Obaidoon and Wei (2024) note that while Copilot excels in addressing grammatical and structural issues, it lacks the capacity for deep thematic analysis or creative evaluation. Similarly, Esfandiari and Allaf-Akbary (2024) observe that Copilot effectively addresses structural and rhetorical aspects of writing. However, more subjective or creative dimensions can be generally better assessed through human judgement to mitigate risks that could be present in AI-generated responses (Robertson et al., 2024).

### *AI-augmented feedback in higher educational contexts*

Traditional feedback practices in higher education often fall short of meeting student expectations, leading to dissatisfaction (Henderson et al., 2019). To address this issue, Schultze et al. (2024) propose AI-augmented feedback as a potential solution, offering a combination of efficiency, scalability, and personalisation while mitigating some of the limitations associated with human-delivered feedback.

AI-augmented feedback provides several advantages that address the challenges of traditional feedback systems. Agostini (2024) highlights how large language models (LLMs) utilise advanced natural language processing to conduct rubric-based assessments, ensuring consistency and reliability in evaluating written tasks across multiple attempts. Similarly, Sol and Heng

(2024) emphasise that AI tools offer personalised academic writing support, particularly for non-native English speakers, by effectively addressing linguistic and structural challenges. Another significant advantage is the enhanced accuracy and efficiency offered by AI-automated grading and feedback systems (Mao et al., 2024). Similar benefits are observed in hybrid intelligent feedback systems, which combine AI-driven automation for technical or repetitive tasks with human judgement to deliver more nuanced interpretations and empathetic responses (Banihashem et al., 2025). AI algorithms can enhance personalisation by providing tailored recommendations and feedback that can suit individual learning needs (Mutambik, 2024). In addition, Shamim et al. (2024) highlight the potential of AI-generated feedback to provide constructive insights on essay assignments, helping students identify areas for improvement during formative assessments. Their study also demonstrates how AI can enhance grading consistency and quality in summative assessments, contributing to the upholding of academic standards.

However, AI tools are not without limitations, particularly when addressing complex or nuanced aspects of writing. Groves et al. (2024) show that Microsoft Copilot excels in technical and structural support but is less effective with tasks requiring creative interpretation or deep contextual understanding. In contrast, Esfandiari and Allaf-Akbary (2024) position Copilot primarily as a form-focused learning tool that supports the identification and practice of international metadiscourse, rather than a tool that provides deeper contextual engagement. Similarly, Misanchuk and Hyzyk (2024) state that ChatGPT, though effective for structured writing, often fails to address the originality and depth required in more nuanced academic tasks. In addition, ethical considerations in the use of AI remain a significant topic of debate. The importance of establishing robust frameworks to protect sensitive academic and personal data has been highlighted, with vulnerabilities in existing systems emphasised (Saif et al., 2024). Patel et al. (2023) point to the risks of data misuse and breaches, advocating for stringent data governance practices to ensure the secure handling of sensitive information. Therefore, educators must understand how AI-derived scores or comments are generated to ensure accountability and prevent misuse (Ejjami, 2024; Griszbacher, 2024).

### *Effective AI* prompts for enhanced AI-augmented feedback

Prompt design plays an important role in the effectiveness of AI-driven educational tools, serving as the interface between users and AI systems. Bozkurt (2023) emphasises the significance of prompt engineering, likening it to crafting well-designed assessment questions that directly influence the quality of AI interactions. This process facilitates personalised learning and enhances problem-solving by eliciting precise and relevant responses from AI systems (Bozkurt, 2023). Effective prompt design often requires an iterative approach, as highlighted by Sikha et al. (2023), who advocate refining prompts through trial-and-error and incorporating user feedback loops to align responses more closely with desired outcomes. Bansal et al. (2024) further reinforce this notion, stressing the importance of specificity and clarity in prompt design to improve feedback effectiveness. Incorporating domain-specific details and structured frameworks can further enhance the quality of prompts. Sikha et al. (2023) suggest that including such elements reduces ambiguity, aligns more effectively with user intentions, and improves the relevance of AI responses. Similarly, Garg and Rajendran (2024) recommend integrating scaffolding and frameworks into prompt design, enabling AI systems to deliver feedback tailored to specific learning objectives and ensuring that responses are both relevant and actionable. Customisation is another vital component of prompt design. Kuzminykh et al. (2024) propose that prompts should be tailored to fit individual user profiles and learning contexts, addressing specific needs more effectively. Their framework suggests that generic prompts often fail to meet particular educational requirements, while personalised prompts significantly enhance the relevance and usefulness of feedback, particularly in individualised learning scenarios.

Despite these opportunities, challenges in prompt design persist. Robertson et al. (2024) note that AI systems may demonstrate hallucinations and biases in their responses. The authors emphasise the importance of iterative refinement, structural clarity, and rigorous human evaluation to mitigate such risks. Similarly, Bansal (2024) underscores the importance of iterative experimentation with various prompt formats to optimise results and help drive innovations in a variety of fields. Additionally, practical strategies such as incorporating detailed examples within prompts, as recommended by Dang et al. (2022), have proven effective in improving AI systems' understanding of users' intentions.

A review of the literature reveals that the field is still in the early, exploratory stages of examining the power and effectiveness of applying AI tools in educational practices, particularly for generating augmented feedback to enhance overall feedback quality. There is also a notable lack of studies conducted in the UK context, especially at the master's level, leaving limited practical guidance on the effective use of AI-augmented feedback. Such a gap in the research prompted the current study, which aims to explore the differences between lecturer-generated feedback and AI-augmented feedback, identify the strengths and limitations of the latter, and propose strategies for optimising input prompts to facilitate the thoughtful integration of AI-augmented feedback in educational settings.

## Methodology

This study selected a total of eight written assignments at the master's level, with original feedback provided by various lecturers from the same department but associated with two distinct taught courses. Table 1 summarises the characteristics of the selected assignments and their corresponding original feedback. The assignments varied in type, including research proposals, essays, and dissertations, ranging from both formative and summative assessments. The original feedback also demonstrated variations in length, format, structure, and levels of specificity.

Table 1: Characteristics of the selected assignments and original feedback

| Item | Field of study | Type of assignment | Mode of assignment | Word count of assignment | Word count of original feedback |
|---|---|---|---|---|---|
| Assignment 1 | Educational Leadership | Research proposal | Summative | 5181 | 120 |
| Assignment 2 | Educational Leadership | Essay | Summative | 5317 | 510 |
| Assignment 3 | Educational Leadership | Essay | Summative | 5498 | 197 |
| Assignment 4 | Educational Leadership | Essay | Summative | 5324 | 181 |
| Assignment 5 | Educational Leadership | Dissertation | Summative | 20373 | 527 |
| Assignment 6 | Education & Technology | Drafted essay | Formative | 2982 | 184 |
| Assignment 7 | Education & Technology | Partially drafted essay | Formative | 555 | 92 |
| Assignment 8 | Education & Technology | Partially drafted essay | Formative | 550 | 117 |

Source: Wang & Galliou (2024)

Two phases of data collection and analysis were conducted. The first phase involved analysing Assignments 1 to 4 as part of the IOE AI & Assessment Challenge in July 2024. Findings from this phase were presented on the final day of the Challenge through an oral presentation. During this phase, the 'precise' model of Copilot – marketed as offering more accurate and advanced responses – was used to generate AI-augmented feedback. However, this model was no longer available when the second phase of data analysis began, which aimed to continue and expand the study by increasing both the sample size and the depth of analysis.

In the second phase, each assignment, along with its original feedback, was uploaded to Copilot, which was subsequently prompted to optimise the feedback across three stages. Phase 2 aimed to explore differences across three scenarios and evaluate the performance of corresponding prompts. In Stage 1, only the original feedback was provided, and Copilot was tasked with optimising it without referencing the written assignment. In Stage 2, both the written assignment and its original feedback were input, allowing Copilot to optimise the feedback with access to the assignment. In Stage 3, the written assignment, original feedback, and grading criteria were all provided, and Copilot was instructed to use the grading criteria as a guide to optimise the feedback based on the assignment. The input prompts used for the three scenarios were as follows:

- Stage 1: Optimise the following feedback.
- Stage 2: Optimise the following feedback with reference to the attached assignment.
- Stage 3: Use the provided grading criteria as a guide to optimise the feedback based on the attached assignment.

The grading criteria used in Stage 3 were identical for both taught courses and focused on three key aspects:

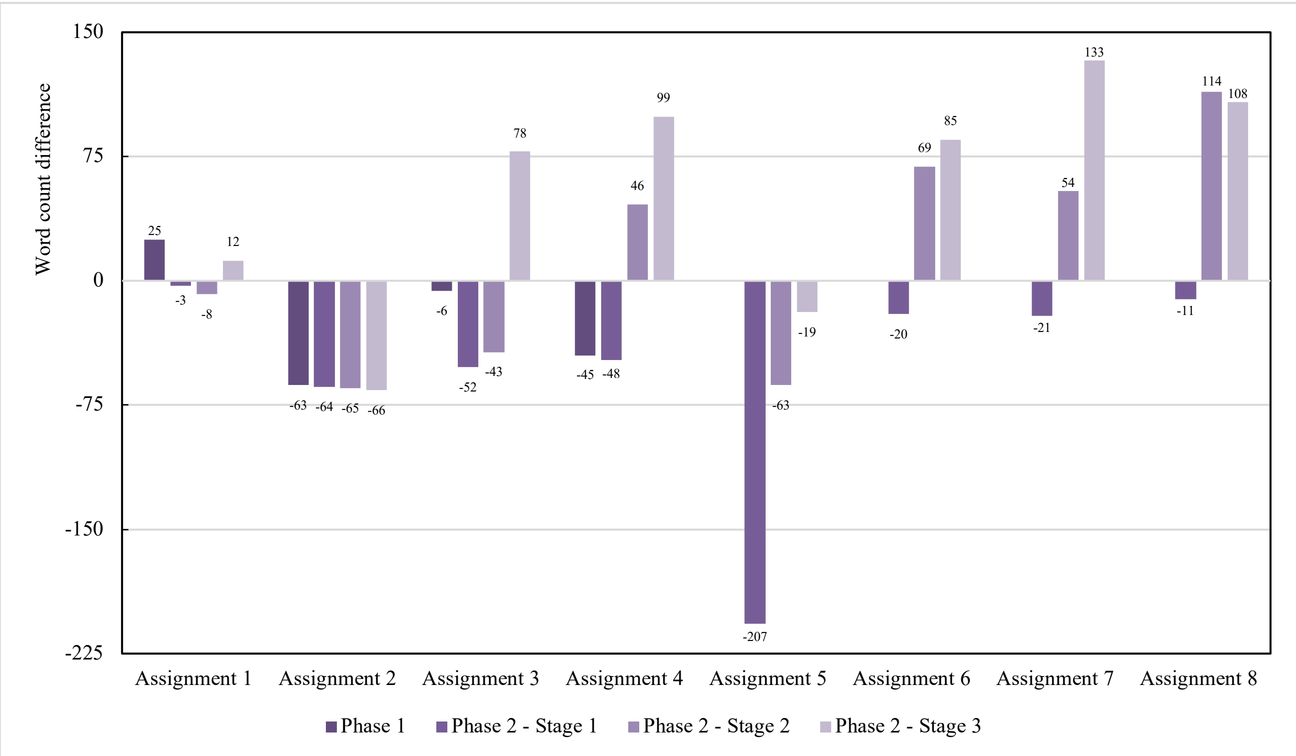
1. Students' grasp of the field of study.
2. Their ability to understand and evaluate research and methodologies.
3. The structure, communication, and presentation of the written assignments.

Following the completion of the three-stage scenarios, a series of follow-up prompts with more specific tasks were entered to further investigate Copilot's responses. To avoid interference with the outcomes generated for different assignments or prompts, a new chat was initiated for each prompt, ensuring that the output remained solely targeted to the specific input. All responses were recorded electronically and analysed to identify key patterns and insights.

## Findings

Initially, the length of the feedback, measured by word count, was assessed and compared. AI-augmented feedback in Phase 1 generally demonstrated smaller and more variable changes compared to the more structured adjustments observed in Phase 2. As shown in Figure 1, AI-augmented feedback was typically shorter than the original feedback in Phase 1 and Stage 1 of Phase 2, as Copilot reduced redundancy and reorganised phrasing to enhance clarity and conciseness. Conversely, in Stages 2 and 3, Copilot increased the length of shorter original feedback to provide more detailed and comprehensive comments. Notably, Stage 2 balanced changes by either reducing or expanding the original content, while Stage 3 presented the most significant positive increases in word count across all assignments, particularly Assignments 3, 6, and 7. This suggests that Copilot intensively added self-generated feedback, focusing on refinement and enhancement during the final stage. However, the significant variability in word count differences across assignments may have been influenced by factors such as the quality, length, or level of specificity of the original feedback.

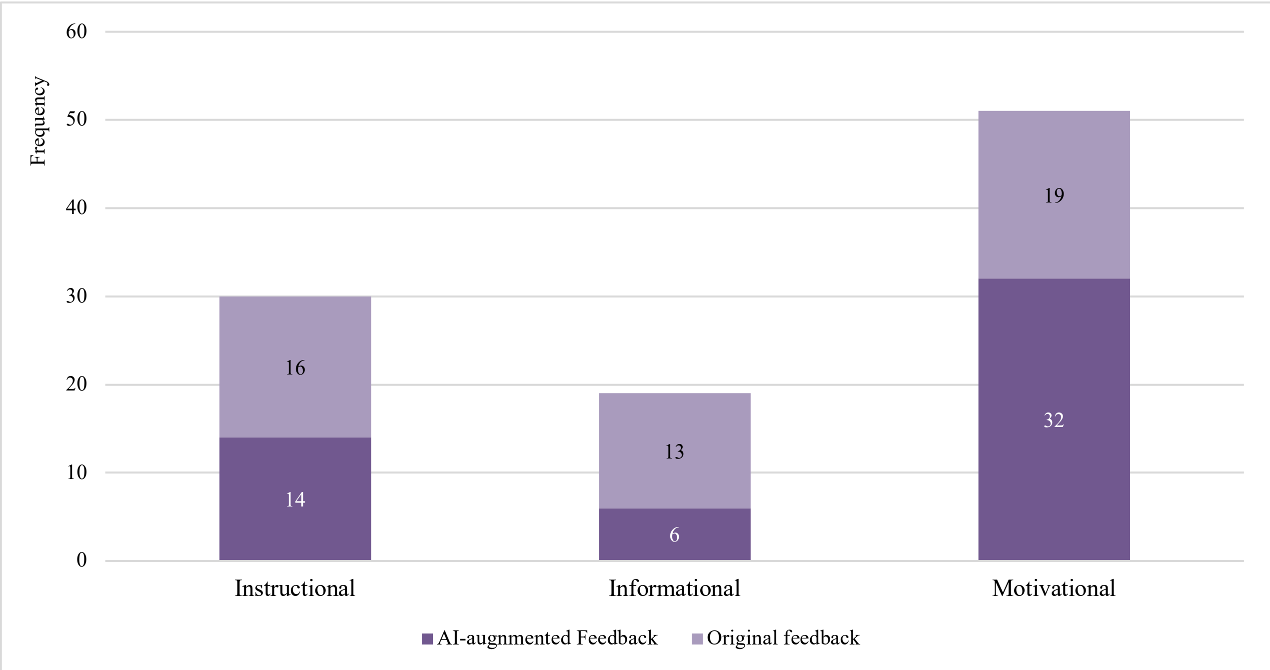
Figure 1: Word count difference of each assignment across two phases




Source: Wang & Galliou (2024)

In the first phase of data analysis, the AI-augmented feedback was segmented into individual sentences, and the probability of each sentence belonging to one of three categories – instructional, informational, or motivational – was determined. Each sentence was assigned to the category with the highest probability, and the frequency of sentences in each category was calculated for comparison. Figure 2 shows that the AI-augmented feedback contained significantly more motivational comments and slightly fewer instructional ones, while the number of informational messages was reduced by Copilot. However, this observation may be subjective and is limited in generalisability due to the small dataset analysed in this phase.

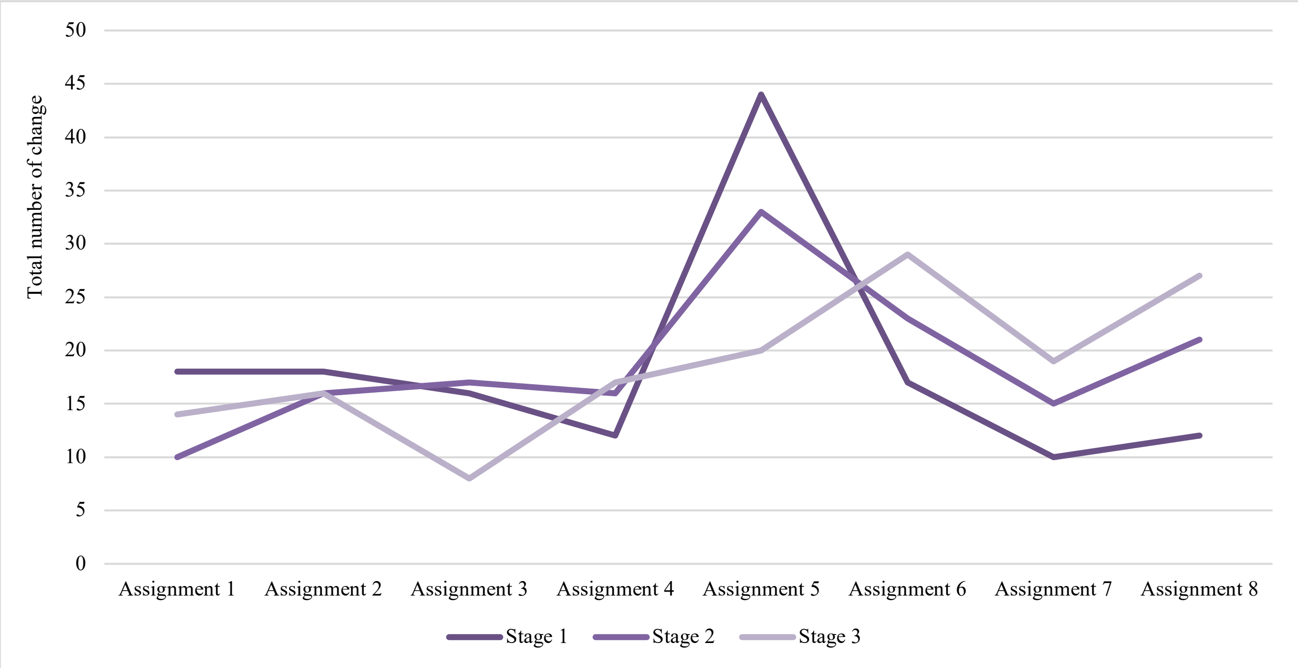
Figure 2: Frequency of feedback categories in orinigal and AI-augmented feedback in Phase 1




Source: Wang & Galliou (2024)

In the second phase, changes in AI-augmented feedback were primarily surface-level in Stage 1, addressing issues such as grammar and formatting. Consequently, the outputs mainly focused on proofreading and providing suggestions for text refinement. In Stage 2, Copilot generated additional content, particularly when the original feedback lacked structured sections. In Stage 3, the amount of additional content increased significantly, especially when the original feedback was brief or did not align with the structured sections of the grading criteria. As a result, data analysis in this phase focused on identifying and coding changes between the original and AI-augmented feedback into four categorised aspects: a) formatting and grammatical changes; b) rewording and rephrasing changes; c) editing and refining changes, and d) additional content. Figure 3 illustrates the total number of changes across the three stages in Phase 2 using a line diagram, while Table 2 explicitly details the number of changes categorised into the four aspects across the three stages for each assignment.

Figure 3: Total number of changes in AI-augmented feedback across the three stages in Phase 2




Source: Wang & Galliou (2024)

As shown in Figure 3, the total number of changes made to feedback in Stage 1 generally started with fewer total changes across assignments, except for Assignment 5, where the number of changes increased significantly to 44. This could be attributed to the longer length of the original feedback, providing more opportunity for minor, initial structural or fundamental changes, such as rewording and rephrasing, to enhance consistency and formality. For other assignments, changes in Stage 1 remained relatively stable, ranging between 10 and 18 changes. In Stage 2, the number of changes increased for most assignments compared to Stage 1, reflecting a tendency toward more detailed adjustments. While Assignments 5 and 6 demonstrated notable peaks, Stage 2 showed a moderate overall increase in changes. Stage 3 consistently recorded the highest or second-highest number of changes across most assignments, particularly for Assignments 6, 7, and 8, which presented a significant rise in total changes. This likely reflects the additional refinement and final adjustments made during this stage, where Copilot focused on improving the feedback's quality and ensuring alignment with the grading criteria.

Table 2: Number and characteristics of changes across stages in Phase 2

| **Item** | **Number and characteristics of changes across stages** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Stage 1* | | | | | *Stage 2* | | | | | *Stage 3* | | | | |
| | *F&G* | *R&R* | *E&R* | *AC* | *Total* | *F&G* | *R&R* | *E&R* | *AC* | *Total* | *F&G* | *R&R* | *E&R* | *AC* | *Total* |
| *A1* | 5 | 8 | 5 | 0 | 18 | 5 | 3 | 2 | 0 | 10 | 6 | 4 | 4 | 0 | 14 |
| *A2* | 8 | 7 | 3 | 0 | 18 | 7 | 6 | 3 | 0 | 16 | 7 | 6 | 3 | 0 | 16 |
| *A3* | 3 | 7 | 6 | 0 | 16 | 3 | 11 | 3 | 0 | 17 | 2 | 1 | 2 | 3 | 8 |
| *A4* | 2 | 6 | 3 | 1 | 22 | 3 | 5 | 3 | 5 | 16 | 2 | 8 | 5 | 2 | 17 |
| *A5* | 6 | 17 | 19 | 2 | 44 | 3 | 23 | 6 | 1 | 33 | 2 | 11 | 7 | 0 | 20 |
| *A6* | 3 | 7 | 5 | 2 | 17 | 3 | 7 | 8 | 5 | 23 | 2 | 12 | 8 | 7 | 29 |
| *A7* | 1 | 6 | 3 | 0 | 10 | 1 | 3 | 2 | 15 | 15 | 2 | 4 | 2 | 11 | 19 |
| *A8* | 2 | 7 | 2 | 1 | 12 | 2 | 6 | 6 | 21 | 21 | 3 | 5 | 4 | 15 | 27 |
| | Sum: 147<br>Min: 10, Max: 44, M: 18.375<br>Mdn: 16.5, Range: 34 | | | | | Sum: 151<br>Min: 10, Max: 33, M: 18.875<br>Mdn: 16.5, Range: 23 | | | | | Sum: 150<br>Min: 8, Max: 29, M: 18.75<br>Mdn: 18, Range: 21 | | | | |

Source: Wang & Galliou (2024)

Stages 1, 2, and 3 showed comparable total changes (147, 151, and 150, respectively), reflecting a consistent effort across all stages. Corrections of punctuation, spelling, and grammar errors, along with improvements in formatting – such as the use of bullet points and the division of long paragraphs into multiple sections aligned with grading criteria – were categorised as formatting and grammatical changes. Changes in this category were generally more frequent for longer original feedback. Rewording and rephrasing changes included alternative word choices, the expansion of incomplete phrases, and the rephrasing of sentences for improved clarity. Copilot often rewrote sentences to avoid the first person and replaced abbreviations to minimise confusion. However, some of the selected word choices were not well-suited to the academic context or commonly used in feedback presentations, highlighting the importance of review and adjustment to minimise the risk of less professional feedback delivery. Editing and refining changes involved removing or adding words and making recommendations to further emphasise or reduce certain parts of the original content. These changes aimed to improve flow, enhance readability, and refine sentences for a higher level of academic professionalism with consistency in tone. Finally, feedback generated and added by Copilot that was not present in the original feedback was categorised as additional content. Such content was often minimal or absent in Stage 1 but increased significantly in Stage 3. Less additional content was added when the original feedback was detailed, structured, and already addressed the grading criteria. Conversely, an increase in additional content was observed for feedback that was brief or unstructured. In these cases, Copilot often provided comments based on the attached assignments, aligning feedback more closely with the grading criteria. Occasionally, parts of the original feedback were removed entirely and replaced with substantial new information to expand the feedback's scope, particularly when the original feedback was extremely brief. However, some of this additional content consisted of broad or generic comments. As a result, manual checks remain essential to evaluate the accuracy and depth of the feedback and to ensure its overall quality and effectiveness.

## Discussion

### *Research question 1: Differences between original and AI-augmented feedback*

When no specific prompts were provided, AI-augmented feedback primarily addressed minor formatting and grammatical errors, as well as rewording and rephrasing, to offer alternative expressions that improved readability and ensured a more consistent flow. Formatting was enhanced through the use of bullet points and the division of content into sections with clearly stated subtitles. Both shortened and expanded sentences were observed, aiming to improve clarity when assignments and grading criteria were provided as references.

The findings reveal a progression in AI-augmented feedback, evolving from surface-level corrections in Stage 1 to more comprehensive revisions in Stage 2 and advanced enhancements in Stage 3. This progression highlights the adaptability of AI systems in customising feedback when guided by iterative prompts or specific grading criteria. Shute's (2008) framework for effective feedback emphasises that actionable, goal-oriented feedback fosters deeper learning and enhances students' self-regulation. The gradual shift observed in this study aligns with this framework, as the feedback became increasingly detailed and goal-specific with the inclusion of clearer instructions, suggesting that iterative prompts and tailored guidance are crucial for maximising the potential of AI systems in generating effective feedback that supports student learning and improvement.

### *Research question 2: Strengths and limitations of using AI-augmented feedback*

AI-augmented feedback demonstrated several strengths in improving the quality of original feedback. Firstly, it provided quick responses, typically within 10 seconds, enabling the efficient generation of augmented feedback. By identifying grammatical errors and refining sentence structure and formatting, AI tools like Copilot enhanced the professionalism and credibility of feedback, particularly when dealing with large volumes of assignments. This aligns with findings from Schultze et al. (2024), which highlight strong student preference for LLM augmented feedback due to its motivating tone and readability, and Misanchuk and Hyzyk (2024), who suggest that AI systems can promote consistency and scalability in STEM education contexts, for example, through structured prompting for reliable outputs.

High consistency was achieved through unified formatting and the systematic organisation of feedback, such as aligning comments with grading criteria. Copilot also effectively clarified unclear phrasing and eliminated repetitive sections, enhancing the readability and usability of feedback. These observations are consistent with studies noting that AI-generated feedback is significantly more readable, detailed, fluent, coherent, and positive (Dai et al., 2023), however less valid in content compared to human feedback. Well-structured feedback, as suggested by Glazzard and Stones (2019), can minimise confusion and help students apply recommendations effectively.

In Phase 1, the use of positive and specific language in AI-augmented feedback enhanced its motivational impact. Supportive and constructive feedback, as noted by Meyer et al. (2024) and Carless (2015), can reduce student anxiety, foster a growth mindset, and improve self-efficacy. Additionally, Copilot's ability to articulate both strengths and weaknesses provided a balanced view of performance, aligning with Hattie and Timperley's (2007) framework for effective feedback. In contrast, Stage 1 of the second phase demonstrated more objective and neutral feedback, with relatively less new content generated. This shift toward unbiased assessments reflects Carless's (2015) perspective on feedback as a relational and dialogic process aimed at guiding rather than directing.

Further, combining task-focused AI feedback with motivational or emotionally supportive elements can be beneficial. For example, Marwan et al. (2022) show that pairing neutral, subgoal-level progress feedback with encouragement supports both performance and persistence, while Alsaiari et al. (2025) demonstrate that emotionally enriched AI feedback can reduce negative emotions and enhance perceived usefulness without compromising feedback actionability, although clarity and relevance require careful design.

Despite these strengths, the absence of the previous Copilot's 'precise' model in the second phase limited its capacity to elaborate on feedback without detailed prompts. This limitation was particularly evident in its inability to provide in-depth critiques or comments essential for discussing complex ideas or addressing creative content. Even in Phase 1, when the 'precise' model was applied, the augmented feedback relied heavily on templated responses, often resulting in generic comments. Without task-specific prompts, the feedback tended to summarise assignments rather than offer constructive and actionable insights. This finding aligns with the caution raised by Misanchuk and Hyzyk (2024), who argue that while AI excels in tasks requiring precision and consistency, it struggles with those necessitating higher-order thinking or deep contextual understanding. Agostini (2024) and Shika et al. (2023) similarly emphasised the importance of clear, task-oriented prompts for eliciting precise and relevant responses from AI systems. When prompts are vague, AI-generated feedback remains surface-level and lacks actionable insights, restricting its effectiveness even when grading criteria are provided. Additionally, while the AI-augmented feedback was generally more formal in phrasing, inappropriate word choices were occasionally observed, making the feedback less suitable for academic contexts. Related limitations were also noted by Chakrabarty et al. (2025) and Nazli et al. (2025) in their respective research on AI-generated text and AI-powered feedback. Furthermore, the default use of American English without explicit prompts necessitated manual checks to ensure alignment with academic conventions.

An updated or enhanced AI model could address these limitations, improving Copilot's capacity to generate contextually appropriate and detailed feedback. By refining its ability to adapt to academic contexts and addressing the need for actionable insights in response to detailed prompts, future iterations of Copilot could further elevate the quality and effectiveness of AI-augmented feedback.

### *Research question 3: Promoting the use of AI-augmented feedback in written assignments*

Given the wide range of factors involved in enhancing the accessibility and usage of AI-augmented feedback, a key strategy to improve its performance is refining the phrasing of prompts to optimise effectiveness for educators. It is clear that there is no universal 'best' prompt, as variations in phrasing and structure can lead to differing responses across AI tools, even when the primary objective remains the same. However, several strategies can be employed to enhance prompts, ensuring more accurate and effective responses. Table 3 provides examples of task-oriented prompts designed to generate AI-assisted feedback, aiming to improve the accuracy and relevance of AI-generated feedback while reducing the likelihood of receiving generic or vague responses. This approach aligns with Garg and Rajendran's (2024) recommendation to incorporate scaffolding and frameworks into prompts, which helps to reduce ambiguity and tailor feedback to specific learning objectives.

Table 3: Suggestions on prompts to improve original feedback

| *Objective* | *Suggested prompts for improving original feedback* |
|---|---|
| *Optimise the feedback* | Can you improve the feedback for better flow and clarity?<br>Can you provide examples to clarify your suggestions?<br>How can I make my feedback more concise? |
| *Further elaboration* | Based on the current feedback, are there any other areas I should comment on?<br>Can you give suggestions on areas I should further elaborate on?<br>Are there any additional points I should consider to enhance my feedback? |
| *Structure and language* | Can you optimise my feedback and proofread it in British English?<br>Can you ensure the texts are presented in a clear structure with a logical flow?<br>Can you simplify complex sentences without twisting meaning? |
| *Engagement and tone* | Can you suggest ways to maintain a supportive tone throughout the feedback?<br>Are there any phrases or words that could be more motivating?<br>How can I balance constructive criticism with positive reinforcements? |

Source: Wang & Galliou (2024)

Refining prompts iteratively allows educators to experiment with phrasing and structures to achieve desired results, incorporating user feedback to enhance the relevance and clarity of AI-augmented feedback. Breaking down complex tasks into distinct instructions and providing examples of ideal outputs can help align feedback with learning outcomes, as highlighted by Molina-Moreira et al. (2023). Additionally, requesting follow-up responses or asking the AI to elaborate on specific areas ensures the feedback effectively addresses strengths and areas for improvement, aligning with Dawson et al.'s (2018) recommendations for providing actionable and meaningful insights.

Furthermore, setting a tone for feedback – such as supportive, constructive, or neutral – can further improve delivery, making AI-augmented feedback more engaging and student-friendly. Meyer et al. (2024) and Carles (2015) highlight that emotionally supportive and constructive feedback fosters a positive learning environment, reducing student anxiety and encouraging greater engagement. Similarly, Dweck (2006) suggests that tailoring the tone and focus of feedback addresses diverse student needs while promoting a growth mindset.

Overall, these findings underscore the crucial role of effective prompt design in maximising the potential of AI tools for generating feedback. As Agostini (2024) suggests, training educators in prompt engineering could empower them to use AI tools more efficiently, ensuring that feedback is actionable, clear, and aligned with educational goals. Ongoing research into optimising prompts could establish standardised best practices across AI platforms, addressing the variability in outputs observed in this study.

### *Limitations of the study*

This study was limited by its small sample size of only eight master's-level written assignments, restricting the generalisability of the findings. Additionally, the cross-sectional design serves as a foundation for future research to investigate students' comprehension of AI-augmented feedback with a larger sample size and to explore its long-term effects on academic performance through potential longitudinal studies. Given the rapid advancements in AI technology, future research could also examine new and updated models, as well as a diverse range of generative AI tools, to compare their effectiveness and unique potential in educational contexts.

## Conclusion

This study explored the use of AI-augmented feedback in written assignments at the master's level, with a focus on identifying differences between lecturer-generated and AI-augmented feedback, discussing the strengths and limitations of the latter, and proposing strategies to optimise the generation of AI feedback. By employing Microsoft Copilot, the research aimed to investigate the role of generative AI in enhancing feedback quality and promoting its practical integration in educational contexts.

The findings suggest that AI-augmented feedback demonstrated a capacity to address surface-level issues such as grammar, formatting, and clarity, while also providing structured improvements when guided by specific prompts. However, without task-specific instructions, the feedback often lacked depth and actionable insights, tending to summarise assignments rather than providing comprehensive critiques. In addition, the study highlighted significant strengths of AI-augmented feedback, including its efficiency, consistency, and use of motivational language, aligning with established practices for effective feedback. Nonetheless, limitations such as reliance on templated responses, occasional inappropriate word choices, and a lack of critical engagement with complex ideas highlighted areas for improvement. Finally, suggestions on emphasising the importance of prompt refinement for improving AI-augmented feedback were proposed.

Nevertheless, the study's findings have several implications for educational practices. By integrating AI-augmented feedback into assessment processes, educators can streamline workflows, ensure consistency, and provide more accessible and personalised support to students. However, these benefits must be balanced with ethical considerations, including transparency, data privacy, and the need to maintain critical human oversight. Training educators in prompt engineering and the thoughtful deployment of AI tools is also essential to maximise their potential while addressing limitations.

Future research could expand on this study by involving larger and more diverse sample sizes to improve the generalisability of findings. Longitudinal studies are also recommended to examine the long-term impact of AI-augmented feedback on student learning outcomes and engagement. Additionally, exploring the effectiveness of newer AI models and generative AI tools across various educational levels and disciplines can provide a more comprehensive understanding of their capabilities and limitations. Finally, developing standardised practices for integrating AI into educational feedback processes can be valuable, though adaptation should be approached with careful consideration of pedagogical goals and institutional policies. Addressing these areas through ongoing research will further refine the use of AI in education, solidifying its role as a supportive tool for enhancing teaching and learning practices.

## Abbreviations

AI: Artificial Intelligence

EFL: English as a Foreign Language

GenAI: Generative Artificial Intelligence

LLM: Large Language Model

NLP: Natural Language Processing

## References


Agostini, D. (2024). Are large language models capable of assessing students' written products? A pilot study in higher education. Research Trends in Humanities, 11(2024), 38–61.

Alsaiari, O., Baghaei, N., Lahza, H., Lodge, J. M., Boden, M., & Khosravi, H. (2025). Emotionally enriched AI-generated feedback: Supporting student well-being without compromising learning. Computers & Education, 239(2025), 1–20. https://doi.org/10.1016/j.compedu.2025.105363

Banihashem, S. K., Noroozi, O., Khosravi, H., Schunn, C. D., & Drachsler, H. (2025). Pedagogical framework for hybrid intelligent feedback. Innovations in Education and Teaching International, 1–17. https://doi.org/10.1080/14703297.2025.2499174

Bansal, P. (2024). Prompt engineering importance and applicability with generative AI. Journal of Computer and Communications, 12(10), 14–23. https://doi.org/10.4236/jcc.2024.1210002

Beaumont, C., O'Doherty, M., & Shannon, L. (2008). Staff and student perceptions of feedback quality in the context of widening participation. Higher Education Academy.

Bozkurt, A. (2023). Unleashing the potential of generative AI, conversational agents and chatbots in educational praxis: A systematic review and bibliometric analysis of GenAI in education. Open Praxis, 15(4), 261–270. https://doi.org/10.55982/openpraxis.15.4.609

Brookhart, S. M. (2017). How to give effective feedback to your students. ASCD.

Carless, D. (2015). Excellence in university assessment: Learning from award-winning practice. Routledge.

Chakrabarty, T., Laban, P., & Wu, C.-S. (2025). Can AI writing be salvaged? Mitigating idiosyncrasies and improving human-AI alignment in the writing process through edits. In Proceedings of the CHI Conference on Human Factors in Computing Systems (pp.1-33). https://doi.org/10.1145/3706598.3713559

Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial Intelligence trends in education: A narrative overview. Procedia computer science, 136, 16–24. https://doi.org/10.1016/j.procs.2018.08.233

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (2023, July). 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

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). How to prompt? Opportunities and challenges of zero- and few-shot learning for human-AI interaction in creative applications of generative models. Generative AI and HCI Workshop at CHI 2022. https://doi.org/10.48550/arXiv.2209.01390

Dawson, P., Henderson, M., Mahoney, P., Phillips, M., Ryan, T., Boud, D., & Molloy, E. (2018). What makes for effective feedback: Staff and student perspectives. Assessment & Evaluation in Higher Education, 44(1), 25–36. https://doi.org/10.1080/02602938.2018.1467877

Dweck, C. (2006). Mindset: The new psychology of success. Random House.

Eager, B., & Brunton, R. (2023). Prompting higher education towards AI-augmented teaching and learning practice. Journal of University Teaching and Learning Practice, 20(5), 1–21. https://doi.org/10.53761/1.20.5.02

Ejjami, R. (2024). The future of learning: AI-based curriculum development. International Journal for Multidisciplinary Research, 6(4), 1–31. https://doi.org/10.36948/ijfmr.2024.v06i04.24441

Esfandiari, R., & Allaf-Akbary, O. (2024). Assessing interactional metadiscourse in EFL writing through intelligent data-driven learning: The Microsoft Copilot in the spotlight. Language Testing in Asia, 14(51), 1–30. https://doi.org/10.1186/s40468-024-00326-9

Garg, A., & Rajendran, R. (2024). The impact of structured prompt-driven generative AI on learning data analysis in engineering students. In Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024), 2, 270–277. https://doi.org/10.5220/0012693000003693

Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leo, D., ... & Rienties, B. (2024). The promise and challenges of generative AI in education. Behaviour & Information Technology, 1–27. https://doi.org/10.1080/0144929X.2024.2394886

Glazzard, J., & Stones, S. (2019). Student perceptions of feedback in higher education. International Journal of Learning, Teaching and Educational Research, 18(11), 38–52. https://doi.org/10.26803/ijlter.18.11.3

Griszbacher, N. (2024). Delving into chit-chat with GPT-3.5: Holy grail or Pandora's Box? A review of AI opportunities and challenges in academia. GILE Journal of Skills Development, 4(1), 4–29. http://doi.org/10.52398/gjsd.2024.v4.i1.pp4-29

Groves, M., Garralda Ortega, Á., Ting, A., & Zou, D. (2024). Using Microsoft Copilot to paraphrase in Eap: An exploratory study of its potentials and pitfalls. SSRN. http://dx.doi.org/10.2139/ssrn.4962522

Guan, C., Mou, J., & Jiang, Z. (2020). Artificial intelligence innovation in education: A twenty-year data-driven historical analysis. International Journal of Innovation Studies, 4(4), 134–147. https://doi.org/10.1016/j.ijis.2020.09.001

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

Hattie, J. A., & Yates, G. C. (2014). Using feedback to promote learning. In V. A. Benassi, C. E. Overson., & C. M. Hakala (Eds.), Applying the science of learning in education: Infusing psychological science into the curriculum (pp. 45–58). American Psychological Association. http://teachpsych.org/ebooks/asle2014/

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

Kuzminykh, I., Nawaz, T., Shenzhang, S., Ghita, B., Raphael, J., & Xiao, H. (2024). Personalised feedback framework for online education programmes using generative AI. https://doi.org/10.48550/arXiv.2410.11904

Lynam, S., & Cachia, M. (2018). Students' perceptions of the role of assessments at higher education. Assessment & Evaluation in Higher Education, 43(2), 223–234. https://doi.org/10.1080/02602938.2017.1329928

Mao, J., Chen, B., & Liu, J. C. (2024). Generative artificial intelligence in education and its implications for assessment. TechTrends, 68(1), 58–66. https://doi.org/10.1007/s11528-023-00911-4

Marwan, S., Akram, B., Barnes, T., & Price, T. W. (2022). Adaptive Immediate Feedback for Block-Based Programming: Design and Evaluation. IEEE Transactions on Learning Technologies, 15(3), 406–420. https://doi.org/10.1109/TLT.2022.3180984

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (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(2024), 1–10. https://doi.org/10.1016/j.caeai.2023.100199

Miao, F., Holmes, W., Huang, R., & Zhang, H. (2021). AI and education: A guidance for policymakers. UNESCO Publishing.

Misanchuk, M., & Hyzyk, J. (2024). ChatGPT in STEM Teaching: An introduction to using LLM-based tools in Higher Ed. https://openlibrary-repo.ecampusontario.ca/jspui/handle/123456789/2121

Molina-Moreira, A. A., Velásquez-Orellana, O. J., Zambrano-Murillo, D. J., & Zambrano-Villamil, M. E. (2023). Importance of feedback in the student evaluation process. International Journal of Social Sciences, 6(3), 168–172. https://doi.org/10.21744/ijss.v6n3.2176

Mutambik, I. (2024). The use of AI-driven automation to enhance student learning experiences in the KSA: An alternative pathway to sustainable education. Sustainability, 16(14), 1–21. https://doi.org/10.3390/su16145970

Nazli, K., Jumani, S., & Masum, R. (2025). The effectiveness of AI-powered corrective feedback on students' writing performance. Pakistan Social Sciences Review, 9(1), 391–403. https://doi.org/10.35484/pssr.2025(9-I)31

Nicol, D. (2010). From monologue to dialogue: Improving written feedback processes in mass higher education. Assessment & Evaluation in Higher Education, 35(5), 501–517. https://doi.org/10.1080/02602931003786559

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. Studies in Higher Education, 31(2), 199–218. https://doi.org/10.1080/03075070600572090

Obaidoon, S., & Wei, H. (2024). ChatGPT, Bard, Bing Chat, and Claude generate feedback for Chinese as foreign language writing: A comparative case study. Future in Educational Research, 2(3), 184–204. https://doi.org/10.1002/fer3.39

Patel, R., Bajaj, P., Kumar, A., Kumari, A., Rai, V., & Kumar, S. (2023, December). ChatGPT in the Classroom: A Comprehensive Review of the Impact of ChatGPT on Modern Education. In 2023 11th International Conference on Intelligent Systems and Embedded Design (ISED) (pp. 1–6). IEEE. https://doi.org/10.1109/ISED59382.2023.10444568

Robertson, J., Ferreira, C., Botha, E., & Oosthuizen, K. (2024). Game changers: A generative AI prompt protocol to enhance human-AI knowledge co-construction. Business Horizons, 67(5), 499–510. https://doi.org/10.1016/j.bushor.2024.04.008

Saif, M., Iqbal, F., Sheaikh, R. A., Ilyas, A., Farid, G., & Nadeem, M. (2024). Ethical considerations regarding the use of artificial intelligence tools in research and publication. Kurdish Studies, 12(4), 1166–1170. https://doi.org/10.53555/ks.v12i4.3133

Schultze, T., Kumar, V. S., McKeown, G., O'Connor, P. A., Rychlowska, M., & Sparemblek, K. (2024). Using Large Language Models to augment (rather than replace) human feedback in higher education improves perceived feedback quality. Queen's University Belfast. https://doi.org/10.31234/osf.io/tvcag

Selwyn, N. (2022). The future of AI and education: Some cautionary notes. European Journal of Education, 57(4), 620–631. https://doi.org/10.1111/ejed.12532

Shamim, M. S., Zaidi, S. J. A., & Rehman, A. (2024). The revival of essay-type questions in medical education: Harnessing artificial intelligence and machine learning. Medical Education, 34(12), 595–599. https://doi.org/10.29271/jcpsp.2024.05.595

Shute, V. J. (2008). Focus on formative feedback. Review of educational research, 78(1), 153–189. https://doi.org/10.3102/0034654307313795

Sikha, V. K., Siramgari, D., & Korada, L. (2023). Mastering prompt engineering: Optimizing interaction with generative AI agents. Journal of Engineering and Applied Sciences Technology, 5(6), 1–8. https://doi.org/10.47363/JEAST/2023(5)E117.

Sol, K., & Heng, K. (2024). AI-powered chatbots as personalized academic writing assistants for non-native English speakers. Encyclopedia of educational innovation.

Spataro, J. (2023, March 16). Introducing Microsoft 365 Copilot - your copilot for work. Official Microsoft Blog. https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/