Banihashem *et al.*
International Journal of Educational Technology in Higher Education (2024) 21:23
https://doi.org/10.1186/s41239-024-00455-4

International Journal of Educational Technology in Higher Education

RESEARCH ARTICLE **Open Access**

Feedback sources in essay writing: peer-generated or AI-generated feedback?

Seyyed Kazem Banihashem[1,2*], Nafiseh Taghizadeh Kerman[3], Omid Noroozi[2], Jewoong Moon[4] and Hendrik Drachsler[1,5]

*Correspondence:
Seyyed Kazem Banihashem
kazem.banihashem@ou.nl
[1] Open Universiteit, Heerlen, The Netherlands
[2] Wageningen University and Research, Wageningen, The Netherlands
[3] Ferdowsi University of Mashhad, Mashhad, Iran
[4] The University of Alabama, Tuscaloosa, USA
[5] DIPE Leibniz Institute, Goethe University, Frankfurt, Germany

Abstract

Peer feedback is introduced as an effective learning strategy, especially in large-size classes where teachers face high workloads. However, for complex tasks such as writing an argumentative essay, without support peers may not provide high-quality feedback since it requires a high level of cognitive processing, critical thinking skills, and a deep understanding of the subject. With the promising developments in Artificial Intelligence (AI), particularly after the emergence of ChatGPT, there is a global argument that whether AI tools can be seen as a new source of feedback or not for complex tasks. The answer to this question is not completely clear yet as there are limited studies and our understanding remains constrained. In this study, we used ChatGPT as a source of feedback for students' argumentative essay writing tasks and we compared the quality of ChatGPT-generated feedback with peer feedback. The participant pool consisted of 74 graduate students from a Dutch university. The study unfolded in two phases: firstly, students' essay data were collected as they composed essays on one of the given topics; subsequently, peer feedback and ChatGPT-generated feedback data were collected through engaging peers in a feedback process and using ChatGPT as a feedback source. Two coding schemes including coding schemes for essay analysis and coding schemes for feedback analysis were used to measure the quality of essays and feedback. Then, a MANOVA analysis was employed to determine any distinctions between the feedback generated by peers and ChatGPT. Additionally, Spearman's correlation was utilized to explore potential links between the essay quality and the feedback generated by peers and ChatGPT. The results showed a significant difference between feedback generated by ChatGPT and peers. While ChatGPT provided more descriptive feedback including information about how the essay is written, peers provided feedback including information about identification of the problem in the essay. The overarching look at the results suggests a potential complementary role for ChatGPT and students in the feedback process. Regarding the relationship between the quality of essays and the quality of the feedback provided by ChatGPT and peers, we found no overall significant relationship. These findings imply that the quality of the essays does not impact both ChatGPT and peer feedback quality. The implications of this study are valuable, shedding light on the prospective use of ChatGPT as a feedback source, particularly for complex tasks like argumentative essay writing.

We discussed the findings and delved into the implications for future research and practical applications in educational contexts.

Keywords AI-generated feedback, ChatGPT, Essay writing, Feedback sources, Higher education, Peer feedback

Introduction

Feedback is acknowledged as one of the most crucial tools for enhancing learning (Banihashem et al., 2022). The general and well-accepted definition of feedback conceptualizes it as information provided by an agent (e.g., teacher, peer, self, AI, technology) regarding aspects of one's performance or understanding (e.g., Hattie & Timplerely, 2007). Feedback serves to heighten students' self-awareness concerning their strengths and areas warranting improvement, through providing actionable steps required to enhance performance (Ramson, 2003). The literature abounds with numerous studies that illuminate the positive impact of feedback on diverse dimensions of students' learning journey including increasing motivation (Amiryousefi & Geld, 2021), fostering active engagement (Zhang & Hyland, 2022), promoting self-regulation and metacognitive skills (Callender et al., 2016; Labuhn et al., 2010), and enriching the depth of learning outcomes (Gan et al., 2021).

Normally, teachers have primarily assumed the role of delivering feedback, providing insights into students' performance on specific tasks or their grasp of particular subjects (Konold et al., 2004). This responsibility has naturally fallen upon teachers owing to their expertise in the subject matter and their competence to offer constructive input (Diezmann & Watters, 2015; Holt-Reynolds, 1999; Valero Haro et al., 2023). However, teachers' role as feedback providers has been challenged in recent years as we have witnessed a growth in class sizes due to the rapid advances in technology and the widespread use of digital technologies that resulted in flexible and accessible education (Shi et al., 2019). The growth in class sizes has translated into an increased workload for teachers, leading to a pertinent predicament. This situation has directly impacted their capacity to provide personalized and timely feedback to each student, a capability that has encountered limitations (Er et al., 2021).

In response to this challenge, various solutions have emerged, among which peer feedback has arisen as a promising alternative instructional approach (Er et al., 2021; Gao et al., 2024; Noroozi et al., 2023; Kerman et al., 2024). Peer feedback entails a process wherein students assume the role of feedback providers instead of teachers (Liu & Carless, 2006). Involving students in feedback can add value to education in several ways. First and foremost, research indicates that students delve into deeper and more effective learning when they take on the role of assessors, critically evaluating and analyzing their peers' assignments (Gielen & De Wever, 2015; Li et al., 2010). Moreover, involving students in the feedback process can augment their self-regulatory awareness, active engagement, and motivation for learning (e.g., Arguedas et al., 2016). Lastly, the incorporation of peer feedback not only holds the potential to significantly alleviate teachers' workload by shifting their responsibilities from feedback provision to the facilitation of peer feedback processes but also nurtures a dynamic learning environment wherein students are actively immersed in the learning journey (e.g., Valero Haro et al., 2023).

Despite the advantages of peer feedback, furnishing high-quality feedback to peers remains a challenge. Several factors contribute to this challenge. Primarily, generating effective feedback necessitates a solid understanding of feedback principles, an element that peers often lack (Latifi et al., 2023; Noroozi et al., 2016). Moreover, offering high-quality feedback is inherently a complex task, demanding substantial cognitive processing to meticulously evaluate peers' assignments, identify issues, and propose constructive remedies (King, 2002; Noroozi et al., 2022). Furthermore, the provision of valuable feedback calls for a significant level of domain-specific expertise, which is not consistently possessed by students (Alqassab et al., 2018; Kerman et al., 2022).

In recent times, advancements in technology, coupled with the emergence of fields like Learning Analytics (LA), have presented promising avenues to elevate feedback practices through the facilitation of scalable, timely, and personalized feedback (Banihashem et al., 2023; Deeva et al., 2021; Drachsler, 2023; Drachsler & Kalz, 2016; Pardo et al., 2019; Zawacki-Richter et al., 2019; Rüdian et al., 2020). Yet, a striking stride forward in the field of educational technology has been the advent of a novel Artificial Intelligence (AI) tool known as "ChatGPT," which has sparked a global discourse on its potential to significantly impact the current education system (Ray, 2023). This tool's introduction has initiated discussions on the considerable ways AI can support educational endeavors (Bond et al., 2024; Darvishi et al., 2024).

In the context of feedback, AI-powered ChatGPT introduces what is referred to as AI-generated feedback (Farrokhnia et al., 2023). While the literature suggests that ChatGPT has the potential to facilitate feedback practices (Dai et al., 2023; Katz et al., 2023), this literature is very limited and mostly not empirical leading us to realize that our current comprehension of its capabilities in this regard is quite restricted. Therefore, we lack a comprehensive understanding of how ChatGPT can effectively support feedback practices and to what degree it can improve the timeliness, impact, and personalization of feedback, which remains notably limited at this time.

More importantly, considering the challenges we raised for peer feedback, the question is whether AI-generated feedback and more specifically feedback provided by ChatGPT has the potential to provide quality feedback. Taking this into account, there is a scarcity of knowledge and research gaps regarding the extent to which AI tools, specifically ChatGPT, can effectively enhance feedback quality compared to traditional peer feedback. Hence, our research aims to investigate the quality of feedback generated by ChatGPT within the context of essay writing and to juxtapose its quality with that of feedback generated by students.

This study carries the potential to make a substantial contribution to the existing body of recent literature on the potential of AI and in particular ChatGPT in education. It can cast a spotlight on the quality of AI-generated feedback in contrast to peer-generated feedback, while also showcasing the viability of AI tools like ChatGPT as effective automated feedback mechanisms. Furthermore, the outcomes of this study could offer insights into mitigating the feedback-related workload experienced by teachers through the intelligent utilization of AI tools (e.g., Banihashem et al., 2022; Er et al., 2021; Pardo et al., 2019).

However, there might be an argument regarding the rationale for conducting this study within the specific context of essay writing. Addressing this potential query, it is crucial to highlight that essay writing stands as one of the most prevalent yet complex tasks for students (Liunokas, 2020). This task is not without its challenges, as evidenced by the extensive body of literature that indicates students often struggle to meet desired standards in their essay composition (e.g., Bulqiyah et al., 2021; Noroozi et al., 2016;, 2022; Latifi et al., 2023).

Furthermore, teachers frequently express dissatisfaction with the depth and overall quality of students' essay writing (Latifi et al., 2023). Often, these teachers lament that their feedback on essays remains superficial due to the substantial time and effort required for critical assessment and individualized feedback provision (Noroozi et al., 2016;, 2022). Regrettably, these constraints prevent them from delving deeper into the evaluation process (Kerman et al., 2022).

Hence, directing attention towards the comparison of peer-generated feedback quality and AI-generated feedback quality within the realm of essay writing bestows substantial value upon both research and practical application. This study enriches the academic discourse and informs practical approaches by delivering insights into the adequacy of feedback quality offered by both peers and AI for the domain of essay writing. This investigation serves as a critical step in determining whether the feedback imparted by peers and AI holds the necessary caliber to enhance the craft of essay writing.

The ramifications of addressing this query are noteworthy. Firstly, it stands to significantly alleviate the workload carried by teachers in the process of essay evaluation. By ascertaining the viability of feedback from peers and AI, teachers can potentially reduce the time and effort expended in reviewing essays. Furthermore, this study has the potential to advance the quality of essay compositions. The collaboration between students providing feedback to peers and the integration of AI-powered feedback tools can foster an environment where essays are not only better evaluated but also refined in their content and structure.With this in mind, we aim to tackle the following key questions within the scope of this study:

RQ1. To what extent does the quality of peer-generated and ChatGPT-generated feedback differ in the context of essay writing?

RQ2. Does a relationship exist between the quality of essay writing performance and the quality of feedback generated by peers and ChatGPT?

Method

Context and participant

This study was conducted in the academic year of 2022–2023 at a Dutch university specializing in life sciences. In total, 74 graduate students from food sciences participated in this study in which 77% of students were female ($N=57$) and 23% were male ($N=17$).

Study design and procedure

This empirical study has an exploratory nature and it was conducted in two phases. An online module called "*Argumentative Essay Writing*" (AEW) was designed to be followed by students within the Brightspace platform. The purpose of the AEW module was to improve students' essay writing skills by engaging them in a peer learning process where students were invited to provide feedback on each other's essays. After designing the module, the study was implemented in two weeks and followed in two phases.

In week one (phase one), students were asked to write an essay on given topics. The topics for the essay were controversial and included "*Scientists with affiliations to the food industry should abstain from participating in risk assessment processes*", "*powdered infant formula must adhere to strict sterility standards*", and "*safe food consumption is* *the responsibility of the consumer*". The given controversial topics were directly related to the course content and students' area of study. Students had time for one week to write their essays individually and submit them to the Brightspace platform.

In week two (phase two), students were randomly invited to provide two sets of written/asynchronous feedback on their peers' submitted essays. We gave a prompt to students to be used for giving feedback (*Please provide feedback to your peer and explain the extent to which she/he has presented/elaborated/justified various elements of an argumentative essay. What are the problems and what are your suggestions to improve each element of the essay? Your feedback must be between 250 and 350 words*). To be able to engage students in the online peer feedback activity, we used the FeedbackFruits app embedded in the Brightspace platform. FeedbackFruits functions as an external educational technology tool seamlessly integrated into Brightspace, aimed at enhancing student engagement via diverse peer collaboration approaches. Among its features are peer feedback, assignment evaluation, skill assessment, automated feedback, interactive videos, dynamic documents, discussion tasks, and engaging presentations (Noroozi et al., 2022). In this research, our focus was on the peer feedback feature of the FeedbackFruits app, which empowers teachers to design tasks that enable students to offer feedback to their peers.

In addition, we used ChatGPT as another feedback source on peers' essays. To be consistent with the criteria for peer feedback, we gave the same feedback prompt question with a minor modification to ChatGPT and asked it to give feedback on the peers' essays (*Please read and provide feedback on the following essay and explain the extent to which she/he has presented/elaborated/justified various elements of an argumentative essay. What are the problems and what are your suggestions to improve each element of the essay? Your feedback must be between 250 and 350 words*).

Following this design, we were able to collect students' essay data, peer feedback data, and feedback data generated by ChatGPT. In the next step, we used two coding schemes to analyze the quality of the essays and feedback generated by peers and ChatGPT.

Measurements

Coding scheme to assess the quality of essay writing

In this study, a coding scheme proposed by Noroozi et al. (2016) was employed to assess students' essay quality. This coding system was constructed based on the key components of high-quality essay composition, encompassing eight elements: introduction pertaining to the subject, taking a clear stance on the subject, presenting arguments in favor of the chosen position, providing justifications for the arguments supporting the position, counter-arguments, justifications for counter-arguments, responses to counter-arguments, and concluding with implications. Each element in the coding system is assigned a score ranging from zero (indicating the lowest quality level) to three (representing the highest quality level). The cumulative scores across all these elements were aggregated to determine the overall quality score of the student's written essays. Two experienced coders in the field of education collaborated to assess the quality of the written essays, and their agreement level was measured at 75% (Cohen's Kappa = 0.75 [95% confidence interval: 0.70–0.81]; z = 25.05; $p < 0.001$), signifying a significant level of consensus between the coders.

Coding scheme to assess the quality of feedback generated by peers and ChatGPT

To assess the quality of feedback provided by both peers and ChatGPT, we employed a coding scheme developed by Noroozi et al. (2022). This coding framework dissects the characteristics of feedback, encompassing three key elements: the affective component, which considers the inclusion of emotional elements such as positive sentiments like praise or compliments, as well as negative emotions such as anger or disappointment; the cognitive component, which includes description (a concise summary of the essay), identification (pinpointing and specifying issues within the essay), and justification (providing explanations and justifications for the identified issues); and the constructive component, which involves offering recommendations, albeit not detailed action plans for further enhancements. Ratings within this coding framework range from zero, indicating poor quality, to two, signifying good quality. The cumulative scores were tallied to determine the overall quality of the feedback provided to the students. In this research, as each essay received feedback from both peers and ChatGPT, we calculated the average score from the two sets of feedback to establish the overall quality score for the feedback received, whether from peers or ChatGPT. The same two evaluators were involved in the assessment. The inter-rater reliability between the evaluators was determined to be 75% (Cohen's Kappa = 0.75 [95% confidence interval: 0.66–0.84]; z = 17.52; $p < 0.001$), showing a significant level of agreement between them.

The logic behind choosing these coding schemes was as follows: Firstly, from a theoretical standpoint, both coding schemes were developed based on robust and well-established theories. The coding scheme for evaluating essay quality draws on Toulmin's argumentation model (1958), a respected framework for essay writing. It encompasses all elements essential for high-quality essay composition and aligns well with the structure of essays assigned in the chosen course for this study. Similarly, the feedback coding scheme is grounded in prominent works on identifying feedback features (e.g., Nelson & Schunn, 2009; Patchan et al., 2016; Wu & Schunn, 2020), enabling the identification of key features of high-quality feedback (Noroozi et al., 2022). Secondly, from a methodological perspective, both coding schemes feature a transparent scoring method, mitigating coder bias and bolstering the tool's credibility.

Analysis

To ensure the data's validity and reliability for statistical analysis, two tests were implemented. Initially, the Levene test assessed group homogeneity, followed by the Kolmogorov-Smirnov test to evaluate data normality. The results confirmed both group homogeneity and data normality. For the first research question, gender was considered as a control variable, and the MANCOVA test was employed to compare the variations in feedback quality between peer feedback and ChatGPT-generated feedback. Addressing the second research question involved using Spearman's correlation to examine the relationships among original argumentative essays, peer feedback, and ChatGPT-generated feedback.

Results

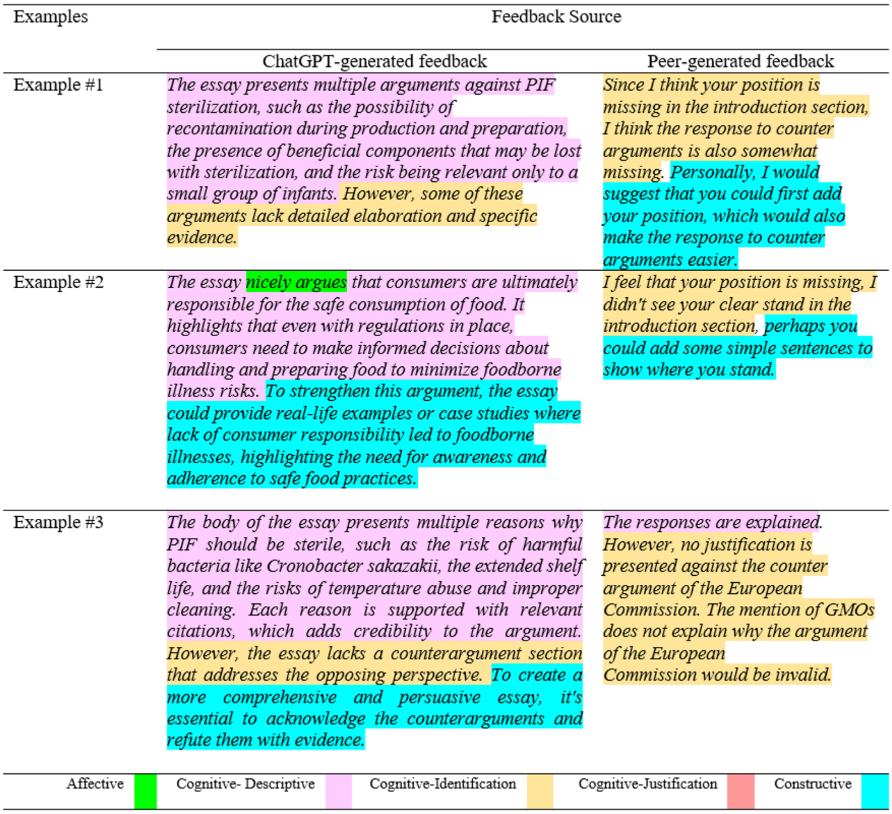
RQ1. To what extent does the quality of peer-generated and ChatGPT-generated feedback differ in the context of essay writing?

The results showed a significant difference in feedback quality between peer feedback and ChatGPT-generated feedback. Peers provided feedback of higher quality compared

Table 1 Differences between peer and ChatGPT-generated feedback in the context of essay writing

Variables		Group	Feedback quality		Difference
			Mean	SD	
Affective		Peer feedback	1.91	0.20	F (1, 146) = 0.32, *p* = 0.48
		ChatGPT feedback	1.93	0.18	
		Total	1.92	0.19	
Cognitive	Description	Peer feedback	1.91	0.29	F (1, 146) = 3.25, *p* < 0.05*, η2 = 0.03
		ChatGPT feedback	2.00	0.00	
		Total	1.95	0.21	
	Identification	Peer feedback	1.52	0.49	F (1, 146) = 4.38, *p* < 0.01**, η2 = 0.02
		ChatGPT feedback	1.29	0.70	
		Total	1.41	0.61	
	Justification	Peer feedback	0.66	0.32	F (1, 146) = 0.24, *p* = 0.36
		ChatGPT feedback	0.62	0.37	
		Total	0.64	0.34	
Constructive		Peer feedback	1.63	0.44	F (1, 146) = 0.36, *p* = 0.26
		ChatGPT feedback	1.68	0.38	
		Total	1.65	0.41	

(*P* < 0.01)**, (*P* < 0.05)*



Examples	Feedback Source	
	ChatGPT-generated feedback	Peer-generated feedback
Example #1	*The essay presents multiple arguments against PIF sterilization, such as the possibility of recontamination during production and preparation, the presence of beneficial components that may be lost with sterilization, and the risk being relevant only to a small group of infants. However, some of these arguments lack detailed elaboration and specific evidence.*	*Since I think your position is missing in the introduction section, I think the response to counter arguments is also somewhat missing. Personally, I would suggest that you could first add your position, which would also make the response to counter arguments easier.*
Example #2	*The essay nicely argues that consumers are ultimately responsible for the safe consumption of food. It highlights that even with regulations in place, consumers need to make informed decisions about handling and preparing food to minimize foodborne illness risks. To strengthen this argument, the essay could provide real-life examples or case studies where lack of consumer responsibility led to foodborne illnesses, highlighting the need for awareness and adherence to safe food practices.*	*I feel that your position is missing, I didn't see your clear stand in the introduction section, perhaps you could add some simple sentences to show where you stand.*
Example #3	*The body of the essay presents multiple reasons why PIF should be sterile, such as the risk of harmful bacteria like Cronobacter sakazakii, the extended shelf life, and the risks of temperature abuse and improper cleaning. Each reason is supported with relevant citations, which adds credibility to the argument. However, the essay lacks a counterargument section that addresses the opposing perspective. To create a more comprehensive and persuasive essay, it's essential to acknowledge the counterarguments and refute them with evidence.*	*The responses are explained. However, no justification is presented against the counter argument of the European Commission. The mention of GMOs does not explain why the argument of the European Commission would be invalid.*

Affective	🟩	Cognitive- Descriptive	🟪	Cognitive-Identification	🟨	Cognitive-Justification	🟥	Constructive	🟦

Fig. 1 A comparative list of selected examples of peer-generated and ChatGPT-generated feedback

Table 2 The relationship between the quality of essays and peer and ChatGPT-generated feedback

Feedback quality		***Essay writing quality***								
		Introduction	***Position***	***Arg. Fav.***	***Just. Fav.***	***Arg. Aga.***	***Just. Aga.***	***Res. Arg.***	***Conclusion***	***Overall***
Affective	*ChatGPT feedback*	0.14	0.19	-0.05	0.09	0.22	0.01	**0.27***	0.14	**0.28***
	Peer feedback	**-0.29***	-0.22	-0.05	-0.15	-0.07	-0.18	-0.04	0.02	**-0.23***
Description	*ChatGPT feedback*	0.12	0.09	0.13	0.09	0.02	0.01	0.04	0.01	0.02
	Peer feedback	**-0.25***	-0.06	-0.11	**-0.23***	0.14	0.00	0.16	-0.13	-0.08
Identification	*ChatGPT feedback*	0.02	-0.16	-0.08	-0.01	-0.14	0.04	-0.15	-0.09	-0.10
	Peer feedback	-0.16	-0.17	0.00	-0.01	0.08	-0.02	0.01	-0.05	-0.06
Justification	*ChatGPT feedback*	0.00	**-0.30***	0.07	0.00	-0.19	-0.03	-0.09	-0.22	-0.18
	Peer feedback	0.01	-0.18	0.06	0.04	0.03	-0.06	0.07	-0.05	-0.02
Constructive	*ChatGPT feedback*	0.04	0.09	-0.19	0.00	0.02	0.05	0.09	-0.09	0.05
	Peer feedback	0.15	-0.16	0.01	0.09	0.10	-0.02	0.10	0.15	0.12
Overall	*ChatGPT feedback*	0.05	-0.16	-0.04	0.01	-0.11	0.02	-0.06	-0.15	-0.08
	Peer feedback	-0.12	**-0.30***	0.00	-0.04	0.11	-0.01	0.12	0.04	-0.05

($P<0.05$)*, ($P<0.01$)**

to ChatGPT. This difference was mainly due to the descriptive and identification of the problem features of feedback. ChatGPT tended to produce more extensive descriptive feedback including a summary statement such as the description of the essay or taken action, while students performed better in pinpointing and identifying the issues in the feedback provided (see Table 1).

A comprehensive list featuring selected examples of feedback generated by peers and ChatGPT is presented in Fig 1. This table additionally outlines examples of how the generated feedback was coded based on the coding scheme to assess the quality of feedback.

RQ2. Does a relationship exist between the quality of essay writing performance and the quality of feedback generated by peers and ChatGPT?

Overall, the results indicated that there was no significant relationship between the quality of essay writing and the feedback generated by peers and ChatGPT. However, a positive correlation was observed between the quality of the essay and the affective feature of feedback generated by ChatGPT, while a negative relationship was observed between the quality of the essay and the affective feature of feedback generated by peers. This finding means that as the quality of the essay improves, ChatGPT tends to provide more affective feedback, while peers tend to provide less affective feedback (see Table 2).

Discussion

This study was an initial effort to explore the potential of ChatGPT as a feedback source in the context of essay writing and to compare the extent to which the quality of feedback generated by ChatGPT differs from the feedback provided by peers. Below we discuss our findings for each research question.

Discussion on the results of RQ1

For the first research question, the results revealed a disparity in feedback quality when comparing peer-generated feedback to feedback generated by ChatGPT. Peer feedback demonstrated higher quality compared to ChatGPT-generated feedback. This discrepancy is attributed primarily to variations in the descriptive and problem-identification features of the feedback.

ChatGPT tended to provide more descriptive feedback, often including elements such as summarizing the content of the essay. This inclination towards descriptive feedback could be related to ChatGPT's capacity to analyze and synthesize textual information effectively. Research on ChatGPT further supports this notion, demonstrating the AI tool's capacity to offer a comprehensive overview of the provided content, therefore potentially providing insights and a holistic perspective on the content (Farrokhnia et al., 2023; Ray, 2023).

ChatGPT's proficiency in providing extensive descriptive feedback could be seen as a strength. It might be particularly valuable for summarizing complex arguments or providing comprehensive overviews, which could aid students in understanding the overall structure and coherence of their essays.

In contrast, students' feedback content entailed high quality regarding identifying specific issues and areas for improvement. Peers outperformance compared to ChatGPT in identifying problems within the essays could be related to humans' potential in cognitive skills, critical thinking abilities, and contextual understanding (e.g., Korteling et al., 2021; Lamb et al., 2019). This means that students, with their contextual knowledge and critical thinking skills, may be better equipped to identify issues within the essays that ChatGPT may overlook.

Furthermore, a detailed look at the findings of the first research question discloses that the feedback generated by ChatGPT comprehensively encompassed all essential components characterizing high-quality feedback, including affective, cognitive, and constructive dimensions (Kerman et al., 2022; Patchan et al., 2016). This comprehensive observation could be an indication of the fact that ChatGPT-generated feedback could potentially serve as a viable source of feedback. This observation is supported by previous studies where a positive role for AI-generated feedback and automated feedback in enhancing educational outcomes has been recognized (e.g., Bellhäuser et al., 2023; Gombert et al., 2024; Huang et al., 2023; Xia et al., 2022).

Finally, an overarching look at the results of the first research question suggests a potential complementary role for ChatGPT and students in the feedback process. This means that using these two feedback sources together creates a synergistic relationship that could result in better feedback outcomes.

Discussion on the results of RQ2

Results for the second research question revealed no observations of a significant correlation between the quality of the essays and the quality of the feedback generated by both peers and ChatGPT. These findings carry a consequential implication, suggesting that the inherent quality of the essays under scrutiny exerts negligible influence over the quality of feedback furnished by both students and the ChatGPT.

In essence, these results point to a notable degree of independence between the writing prowess exhibited in the essays and the efficacy of the feedback received from either source. This disassociation implies that the ability to produce high-quality essays does not inherently translate into a corresponding ability to provide equally insightful feedback, neither for peers nor for ChatGPT. This decoupling of essay quality from feedback quality highlighted the multifaceted nature of these evaluative processes, where proficiency in constructing a coherent essay does not necessarily guarantee an equally adept capacity for evaluating and articulating constructive commentary on peers' work.

The implications of these findings are both intriguing and defy conventional expectations, as they deviate somewhat from the prevailing literature's stance. The existing body of scholarly work generally posits a direct relationship between the quality of an essay and the subsequent quality of generated feedback (Noroozi et al., 2016;, 2022; Kerman et al., 2022; Vale Haro et al., 2023). This line of thought contends that essays of inferior quality might serve as a catalyst for more pronounced error detection among students, encompassing grammatical intricacies, depth of content, clarity, and coherence, as well as the application of evidence and support. Conversely, when essays are skillfully crafted, the act of pinpointing areas for enhancement becomes a more complex task, potentially necessitating a heightened level of subject comprehension and nuanced evaluation.

However, the present study's findings challenge this conventional wisdom. The observed decoupling of essay quality from feedback quality suggests a more nuanced interplay between the two facets of assessment. Rather than adhering to the anticipated pattern, wherein weaker essays prompt clearer identification of deficiencies, and superior essays potentially render the feedback process more challenging, the study suggests that the process might be more complex than previously thought. It hints at a dynamic in which the act of evaluating essays and providing constructive feedback transcends a simple linear connection with essay quality.

These findings, while potentially unexpected, are an indication of the complex nature of essay assignments and feedback provision highlighting the complexity of cognitive processes that underlie both tasks, and suggesting that the relationship between essay quality and feedback quality is not purely linear but influenced by a multitude of factors, including the evaluator's cognitive framework, familiarity with the subject matter, and critical analysis skills.

Despite this general observation, a closer examination of the affective features within the feedback reveals a different pattern. The positive correlation between essay quality and the affective features present in ChatGPT-generated feedback could be related to ChatGPT's capacity to recognize and appreciate students' good work. As the quality of the essay increases, ChatGPT might be programmed to offer more positive and motivational feedback to acknowledge students' progress (e.g., Farrokhnia et al., 2023; Ray, 2023). In contrast, the negative relationship between essay quality and the affective features in peer feedback may be attributed to the evolving nature of feedback from peers (e.g., Patchan et al., 2016). This suggests that as students witness improvements in their peers' essay-writing skills and knowledge, their feedback priorities may naturally evolve. For instance, students may transition from emphasizing emotional and affective comments to focusing on cognitive and constructive feedback, with the goal of further enhancing the overall quality of the essays.

Limitations and implications for future research and practice

We acknowledge the limitations of this study. Primarily, the data underpinning this investigation was drawn exclusively from a singular institution and a solitary course, featuring a relatively modest participant pool. This confined scope inevitably introduces certain constraints that need to be taken into consideration when interpreting the study's outcomes and generalizing them to broader educational contexts. Under this constrained sampling, the findings might exhibit a degree of contextual specificity, potentially limiting their applicability to diverse institutional settings and courses with distinct curricular foci. The diverse array of academic environments, student demographics, and subject matter variations existing across educational institutions could potentially yield divergent patterns of results. Therefore, while the current study's outcomes provide insights within the confines of the studied institution and course, they should be interpreted and generalized with prudence. Recognizing these limitations, for future studies, we recommend considering a large-scale participant pool with a diverse range of variables, including individuals from various programs and demographics. This approach would enrich the depth and breadth of understanding in this domain, fostering a more comprehensive comprehension of the complex dynamics at play.

In addition, this study omitted an exploration into the degree to which students utilize feedback provided by peers and ChatGPT. That is to say that we did not investigate the effects of such feedback on essay enhancements in the revision phase. This omission inherently introduces a dimension of uncertainty and places a constraint on the study's holistic understanding of the feedback loop. By not addressing these aspects, the study's insights are somewhat partial, limiting the comprehensive grasp of the potential influences that these varied feedback sources wield on students' writing enhancement processes. An analysis of the feedback assimilation patterns and their subsequent effects on essay refinement would have unveiled insights into the practical utility and impact of the feedback generated by peers and ChatGPT.

To address this limitation, future investigations could be structured to encompass a more thorough examination of students' feedback utilization strategies and the resulting implications for the essay revision process. By shedding light on the complex interconnection between feedback reception, its integration into the revision process, and the ultimate outcomes in terms of essay improvement, a more comprehensive understanding of the dynamics involved could be attained.

Furthermore, in this study, we employed identical question prompts for both peers and ChatGPT. However, there is evidence indicating that ChatGPT is sensitive to how prompts are presented to it (e.g., Cao et al., 2023; White et al., 2023; Zuccon & Koopman, 2023). This suggests that variations in the wording, structure, or context of prompts might influence the responses generated by ChatGPT, potentially impacting the comparability of its outputs with those of peers. Therefore, it is essential to carefully consider and control for prompt-related factors in future research when assessing ChatGPT's performance and capabilities in various tasks and contexts.

In addition, We acknowledge that ChatGPT can potentially generate inaccurate results. Nevertheless, in the context of this study, our examination of the results generated by ChatGPT did not reveal a significant inaccuracies that would warrant inclusion in our findings.

From a methodological perspective, we reported the interrater reliability between the coders to be 75%. While this level of agreement was statistically significant, signifying the reliability of our coders' analyses, it did not reach the desired level of precision. We acknowledge this as a limitation of the study and suggest enhancing interrater reliability through additional coder training.

In addition, it is worth noting that the advancement of Generative AI like ChatGPT, opens new avenues in educational feedback mechanisms. Beyond just generating feedback, these AI models have the potential to redefine how feedback is presented and assimilated. In the realm of research on adaptive learning systems, the findings of this study also echo the importance of adaptive learning support empowered by AI and ChatGPT (Rummel et al., 2016). It can pave the way for tailored educational experiences that respond dynamically to individual student needs. This is not just about the feedback's content but its delivery, timing, and adaptability. Further exploratory data analyses, such as sequential analysis and data mining, may offer insights into the nuanced ways different adaptive learning supports can foster student discussions (Papamitsiou & Economides, 2014). This involves dissecting the feedback dynamics, understanding how varied feedback types stimulate discourse, and identifying patterns that lead to enhanced student engagement.

Ensuring the reliability and validity of AI-empowered feedback is also crucial. The goal is to ascertain that technology-empowered learning support genuinely enhances students' learning process in a consistent and unbiased manner. Given ChatGPT's complex nature of generating varied responses based on myriad prompts, the call for enhancing methodological rigor through future validation studies becomes both timely and essential. For example, in-depth prompt validation and blind feedback assessment studies could be employed to meticulously probe the consistency and quality of ChatGPT's responses. Also, comparative analysis with different AI models can be useful.

From an educational standpoint, our research findings advocate for the integration of ChatGPT as a feedback resource with peer feedback within higher education environments for essay writing tasks since there is a complementary role potential for pee-generated and ChatGPT-generated feedback. This approach holds the potential to alleviate the workload burden on teachers, particularly in the context of online courses with a significant number of students.

Conclusion

This study contributes to and adds value to the young existing but rapidly growing literature in two distinct ways. From a research perspective, this study addresses a significant void in the current literature by responding to the lack of research on AI-generated feedback for complex tasks like essay writing in higher education. The research bridges this gap by analyzing the effectiveness of ChatGPT-generated feedback compared to peer-generated feedback, thereby establishing a foundation for further exploration in this field. From a practical perspective of higher education, the study's findings offer insights into the potential integration of ChatGPT as a feedback source within higher education contexts. The discovery that ChatGPT's feedback quality could potentially complement peer feedback highlights its applicability for enhancing feedback practices in higher education. This holds particular promise for courses with substantial enrolments and essay-writing components, providing teachers with a feasible alternative for delivering constructive feedback to a larger number of students.

Author contributions

S. K. Banihashem led this research experiment. N. T. Kerman contributed to the data analysis and writing. O. Noroozi contributed to the designing, writing, and reviewing the manuscript. J. Moon contributed to the writing and revising the manuscript. H. Drachsler contributed to the writing and revising the manuscript.

Funding

No funding has been received for this research.

Data availability

The data is available upon a reasonable request.

Declarations

Declaration of AI-assisted technologies in the writing process

The authors used generative AI for language editing and took full responsibility.

Competing interests

The authors declare that they have no competing interests.

Received: 20 November 2023 / Accepted: 18 March 2024

Published online: 12 April 2024

References

Alqassab, M., Strijbos, J. W., & Ufer, S. (2018). Training peer-feedback skills on geometric construction tasks: Role of domain knowledge and peer-feedback levels. *European Journal of Psychology of Education*, *33*(1), 11–30. https://doi.org/10.1007/s10212-017-0342-0.

Amiryousefi, M., & Geld, R. (2021). The role of redressing teachers' instructional feedback interventions in EFL learners' motivation and achievement in distance education. *Innovation in Language Learning and Teaching*, *15*(1), 13–25. https://doi.org/10.1080/17501229.2019.1654482.

Arguedas, M., Daradoumis, A., & Xhafa Xhafa, F. (2016). Analyzing how emotion awareness influences students' motivation, engagement, self-regulation and learning outcome. *Educational Technology and Society*, *19*(2), 87–103. https://www.jstor.org/stable/jeductechsoci.19.2.87.

Banihashem, S. K., Noroozi, O., van Ginkel, S., Macfadyen, L. P., & Biemans, H. J. (2022). A systematic review of the role of learning analytics in enhancing feedback practices in higher education. *Educational Research Review*, 100489. https://doi.org/10.1016/j.edurev.2022.100489.

Banihashem, S. K., Dehghanzadeh, H., Clark, D., Noroozi, O., & Biemans, H. J. (2023). Learning analytics for online game-based learning: A systematic literature review. *Behaviour & Information Technology*, 1–28. https://doi.org/10.1080/0144929X.2023.2255301.

Bellhäuser, H., Dignath, C., & Theobald, M. (2023). Daily automated feedback enhances self-regulated learning: A longitudinal randomized field experiment. *Frontiers in Psychology*, *14*, 1125873. https://doi.org/10.3389/fpsyg.2023.1125873.

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, *21*(4), 1–41. https://doi.org/10.1186/s41239-023-00436-z.

Bulqiyah, S., Mahbub, M., & Nugraheni, D. A. (2021). Investigating writing difficulties in Essay writing: Tertiary Students' perspectives. *English Language Teaching Educational Journal*, *4*(1), 61–73. https://doi.org/10.12928/eltej.v4i1.2371.

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning*, *11*(2), 215–235. https://doi.org/10.1007/s11409-015-9142-6.

Cao, J., Li, M., Wen, M., & Cheung, S. C. (2023). A study on prompt design, advantages and limitations of chatgpt for deep learning program repair. *arXiv Preprint arXiv:2304 08191*. https://doi.org/10.48550/arXiv.2304.08191.

Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y. S., Gasevic, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. https://doi.org/10.35542/osf.io/hcgzj.

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

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, *162*, 104094. https://doi.org/10.1016/j.compedu.2020.104094.

Diezmann, C. M., & Watters, J. J. (2015). The knowledge base of subject matter experts in teaching: A case study of a professional scientist as a beginning teacher. *International Journal of Science and Mathematics Education*, *13*, 1517–1537. https://doi.org/10.1007/s10763-014-9561-x.

Drachsler, H. (2023). *Towards highly informative learning analytics*. Open Universiteit. https://doi.org/10.25656/01:26787.

Drachsler, H., & Kalz, M. (2016). The MOOC and learning analytics innovation cycle (MOLAC): A reflective summary of ongoing research and its challenges. *Journal of Computer Assisted Learning*, *32*(3), 281–290. https://doi.org/10.1111/jcal.12135.

Er, E., Dimitriadis, Y., & Gašević, D. (2021). Collaborative peer feedback and learning analytics: Theory-oriented design for supporting class-wide interventions. *Assessment & Evaluation in Higher Education*, *46*(2), 169–190. https://doi.org/10.1080/02602938.2020.1764490.

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, 1–15. https://doi.org/10.1080/14703297.2023.2195846.

Gan, Z., An, Z., & Liu, F. (2021). Teacher feedback practices, student feedback motivation, and feedback behavior: How are they associated with learning outcomes? *Frontiers in Psychology*, *12*, 697045. https://doi.org/10.3389/fpsyg.2021.697045.

Gao, X., Noroozi, O., Gulikers, J. T. M., Biemans, H. J., & Banihashem, S. K. (2024). A systematic review of the key components of online peer feedback practices in higher education. *Educational Research Review*, 100588. https://doi.org/10.1016/j.edurev.2023.100588.

Gielen, M., & De Wever, B. (2015). Scripting the role of assessor and assessee in peer assessment in a wiki environment: Impact on peer feedback quality and product improvement. *Computers & Education*, *88*, 370–386. https://doi.org/10.1016/j.compedu.2015.07.012.

Gombert, S., Fink, A., Giorgashvili, T., Jivet, I., Di Mitri, D., Yau, J., & Drachsler, H. (2024). From the Automated Assessment of Student Essay Content to highly informative feedback: A case study. *International Journal of Artificial Intelligence in Education*, 1–39. https://doi.org/10.1007/s40593-023-00387-6.

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

Holt-Reynolds, D. (1999). Good readers, good teachers? Subject matter expertise as a challenge in learning to teach. *Harvard Educational Review*, *69*(1), 29–51. https://doi.org/10.17763/haer.69.1.pl5m5083286l77t2.

Huang, A. Y., Lu, O. H., & Yang, S. J. (2023). Effects of artificial intelligence–enabled personalized recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom. *Computers & Education*, *194*, 104684. https://doi.org/10.1016/j.compedu.2022.104684.

Katz, A., Wei, S., Nanda, G., Brinton, C., & Ohland, M. (2023). Exploring the efficacy of ChatGPT in analyzing Student Teamwork Feedback with an existing taxonomy. *arXiv Preprint arXiv*. https://doi.org/10.48550/arXiv.2305.11882.

Kerman, N. T., Noroozi, O., Banihashem, S. K., Karami, M., & Biemans, H. J. (2022). Online peer feedback patterns of success and failure in argumentative essay writing. *Interactive Learning Environments*, 1–13. https://doi.org/10.1080/10494820.2022.2093914.

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

King, A. (2002). Structuring peer interaction to promote high-level cognitive processing. *Theory into Practice*, *41*(1), 33–39. https://doi.org/10.1207/s15430421tip4101_6.

Konold, K. E., Miller, S. P., & Konold, K. B. (2004). Using teacher feedback to enhance student learning. *Teaching Exceptional Children*, *36*(6), 64–69. https://doi.org/10.1177/004005990403600608.

Korteling, J. H., van de Boer-Visschedijk, G. C., Blankendaal, R. A., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human-versus artificial intelligence. *Frontiers in Artificial Intelligence*, *4*, 622364. https://doi.org/10.3389/frai.2021.622364.

Labuhn, A. S., Zimmerman, B. J., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards. *Metacognition and Learning*, *5*, 173–194. https://doi.org/10.1007/s11409-010-9056-2.

Lamb, R., Firestone, J., Schmitter-Edgecombe, M., & Hand, B. (2019). A computational model of student cognitive processes while solving a critical thinking problem in science. *The Journal of Educational Research*, *112*(2), 243–254. https://doi.org/10.1080/00220671.2018.1514357.

Latifi, S., Noroozi, O., & Talaee, E. (2023). Worked example or scripting? Fostering students' online argumentative peer feedback, essay writing and learning. *Interactive Learning Environments*, *31*(2), 655–669. https://doi.org/10.1080/10494820.2020.1799032.

Li, L., & Liu, X. (2010). Steckelberg. Assessor or assessee: How student learning improves by giving and receiving peer feedback. *British Journal of Educational Technology*, *41*(3), 525–536. https://doi.org/10.1111/j.1467-8535.2009.00968.x.

Liu, N. F., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education*, *11*(3), 279–290. https://doi.org/10.1080/13562510600680582.

Liunokas, Y. (2020). Assessing students' ability in writing argumentative essay at an Indonesian senior high school. IDEAS: Journal on English language teaching and learning. *Linguistics and Literature*, *8*(1), 184–196. https://doi.org/10.24256/ideas.v8i1.1344.

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science*, *37*, 375–401. https://doi.org/10.1007/s11251-008-9053-x.

Noroozi, O., Banihashem, S. K., Taghizadeh Kerman, N., Parvaneh Akhteh Khaneh, M., Babayi, M., Ashrafi, H., & Biemans, H. J. (2022). Gender differences in students' argumentative essay writing, peer review performance and uptake in online learning environments. *Interactive Learning Environments*, 1–15. https://doi.org/10.1080/10494820.2022.2034887.

Noroozi, O., Biemans, H., & Mulder, M. (2016). Relations between scripted online peer feedback processes and quality of written argumentative essay. *The Internet and Higher Education*, 31, 20-31. https://doi.org/10.1016/j.iheduc.2016.05.002

Noroozi, O., Banihashem, S. K., Biemans, H. J., Smits, M., Vervoort, M. T., & Verbaan, C. L. (2023). Design, implementation, and evaluation of an online supported peer feedback module to enhance students' argumentative essay quality. *Education and Information Technologies*, 1–28. https://doi.org/10.1007/s10639-023-11683-y.

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, *17*(4), 49–64. https://doi.org/10.2307/jeductechsoci.17.4.49. https://www.jstor.org/stable/.

Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., & Mirriahi, N. (2019). Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*, *50*(1), 128–138. https://doi.org/10.1111/bjet.12592.

Patchan, M. M., Schunn, C. D., & Correnti, R. J. (2016). The nature of feedback: How peer feedback features affect students' implementation rate and quality of revisions. *Journal of Educational Psychology*, *108*(8), 1098. https://doi.org/10.1037/edu0000103.

Ramsden, P. (2003). *Learning to teach in higher education*. Routledge.

Ray, P. P. (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. https://doi.org/10.1016/j.iotcps.2023.04.003.

Rüdian, S., Heuts, A., & Pinkwart, N. (2020). Educational Text Summarizer: Which sentences are worth asking for? *In DELFI 2020 - The 18th Conference on Educational Technologies of the German Informatics Society* (pp. 277–288). Bonn, Germany.

Rummel, N., Walker, E., & Aleven, V. (2016). Different futures of adaptive collaborative learning support. *International Journal of Artificial Intelligence in Education*, *26*, 784–795. https://doi.org/10.1007/s40593-016-0102-3.

Shi, M. (2019). The effects of class size and instructional technology on student learning performance. *The International Journal of Management Education*, *17*(1), 130–138. https://doi.org/10.1016/j.ijme.2019.01.004.

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.

Valero Haro, A., Noroozi, O., Biemans, H. J., Mulder, M., & Banihashem, S. K. (2023). How does the type of online peer feedback influence feedback quality, argumentative essay writing quality, and domain-specific learning? *Interactive Learning Environments*, 1–20. https://doi.org/10.1080/10494820.2023.2215822.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*. https://doi.org/10.48550/arXiv.2302.11382.

Wu, Y., & Schunn, C. D. (2020). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology*, *60*, 101826. https://doi.org/10.1016/j.cedpsych.2019.101826.

Xia, Q., Chiu, T. K., Zhou, X., Chai, C. S., & Cheng, M. (2022). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 100118. https://doi.org/10.1016/j.caeai.2022.100118.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 1–27. https://doi.org/10.1186/s41239-019-0171-0.

Zhang, Z. V., & Hyland, K. (2022). Fostering student engagement with feedback: An integrated approach. *Assessing Writing*, *51*, 100586. https://doi.org/10.1016/j.asw.2021.100586.

Zuccon, G., & Koopman, B. (2023). Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness. *arXiv preprint arXiv:2302*.13793. https://doi.org/10.48550/arXiv.2302.13793.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.