Computers and Education Open 9 (2025) 100249

Contents lists available at ScienceDirect

Computers and Education Open

journal homepage: www.sciencedirect.com/journal/computers-and-education-open

The impact of generative AI on essay revisions and student engagement

Noble Lo [a,b,*], Alan Wong [c], Sumie Chan [c]

[a] Department of Educational Research, Educational Research, County South, Lancaster University, Lancaster, United Kingdom
[b] Division of Languages and Communication, College of Professional and Continuing Education, The Hong Kong Polytechnic University, PolyU West Kowloon Campus, 9 Hoi Ting Road, Yau Ma Tei, Kowloon, Hong Kong, China
[c] English Language Teaching Unit, The Chinese University of Hong Kong, Li Dak Sum Building, Shatin, The New Territories, Hong Kong, China

A R T I C L E I N F O

Keywords:
Generative AI
Feedback
Student engagement
Writing proficiency
AI in education

A B S T R A C T

This study investigates the effect of AI-generated feedback on university students' essay writing proficiency in Hong Kong. By integrating generative AI into the revision process, the research examines how automated feedback influences the quality of students' written work, as well as their engagement, motivation, and emotional responses during revision. The study employs a randomized controlled design, comparing students who received AI feedback on their essays with those who did not. Both quantitative and qualitative data were collected to assess the impact of AI feedback on writing improvement and student experiences. Quantitative analysis shows significant improvements in essay quality for students who utilized AI feedback, while qualitative findings highlight increased engagement and motivation, though students exhibited mixed emotional responses to the revision process. The results suggest that generative AI has considerable potential to enhance writing skills in higher education by providing timely, individualized feedback. This research contributes to ongoing discussions on the role of AI in education, particularly in supporting language instruction and student development.

Practitioner Notes

What is already known about this topic:

Artificial Intelligence (AI) in Education (AIEd) has the potential to provide personalized feedback and support large-scale teaching environments.

Large language models (LLMs) like GPT-3.5 have shown promise in generating coherent and useful feedback for student writing.

Feedback is critical for improving student learning outcomes and engagement, but providing it is time-consuming for educators.

What this paper adds:

This study demonstrates that LLM-based feedback can significantly improve the quality of university students' essay revisions in Hong Kong.

AI-generated feedback enhances student engagement and motivation during the writing revision process.

The emotional experiences of students receiving AI feedback are mixed, indicating both positive and negative reactions to the feedback process.

Implications for practice and policy:

Incorporating LLM-based feedback tools in university-level language courses can improve essay quality and student engagement with writing tasks.

Educators should be aware of the mixed emotional responses to AI feedback and provide additional support to mitigate any negative experiences.

Policies should consider integrating AI tools to reduce teacher workload while maintaining high-quality feedback for students, enhancing overall educational outcomes.

Introduction

The evolution of teaching performance is imperative in response to technological advancements, as education consistently adapts to the integration of emerging technologies [1]. Digital transformation has fundamentally reshaped the approaches to teaching and learning on a global scale [2]. On 20 January 2025, the Hangzhou-based company

* Corresponding author at: Department of Educational Research, Educational Research, County South, Lancaster University, Lancaster, United Kingdom.
E-mail address: p.k.lo@lancaster.ac.uk (N. Lo).

https://doi.org/10.1016/j.caeo.2025.100249
Received 15 July 2024; Received in revised form 31 January 2025; Accepted 13 March 2025
Available online 29 March 2025

DeepSeek launched DeepSeek-R1, a partially open-source large language model (LLM) capable of solving scientific problems at a level comparable to OpenAI's o1, which was introduced in late 2024 [3]. This reasoning model follows a step-by-step process similar to human reasoning, making it both more effective than earlier LLMs and potentially valuable for scientific research [4]. Generative AI, including OpenAI's GPT-4 language model, offers significant potential in education by enabling the creation of personalized learning materials, providing automated and comprehensive feedback to students, and streamlining the grading process for greater efficiency [5]. ChatGPT offers considerable advantages, including assisting educators in generating instructional materials, fostering discussions, and providing personalized feedback to students [6]. This study investigates the use of AI-generated feedback to enhance university students' writing proficiency in Hong Kong, focusing on how it impacts both the quality of their revisions and their levels of engagement, motivation, and emotional responses.

Despite the potential of AI to support writing, delivering high-quality feedback to students remains a considerable challenge. First, as an AI-driven tool, ChatGPT generates inconsistent feedback with each evaluation of student writing, which undermines the reliability and consistency of its assessments [7]. Second, while AI-generated feedback has been shown to improve writing scores in comparison to human feedback, it carries the risk of homogenizing students' writing styles [8]. Lastly, the factual accuracy of AI feedback can be inconsistent, potentially leading to misleading guidance for students [9]. Effectively leveraging AI feedback, therefore, necessitates that students develop critical thinking and digital literacy skills to evaluate the quality and relevance of the feedback provided [10]. However, empirical evidence on the efficacy of generative AI in improving student writing outcomes, particularly in the context of English as a Second Language (ESL) education, is still scarce.

This research is unique in its focus on the application of generative AI in university-level English-language education in Hong Kong, a context not extensively studied in the existing literature. The motivational and affective functions of feedback [11–13] have often been neglected in the context of AI-assisted feedback. This area of research faces several challenges, including a limited number of empirically validated strategies for directly influencing learner affect [14,15]. Furthermore, there exists a significant disparity between the level of "humanness" required for effective affective-motivational communication and the capabilities of AI systems (e.g., [16,17]). This study addresses critical gaps by investigating both the measurable impact of AI-generated feedback on essay revisions and its effect on students' motivational and emotional states during the revision process. The findings contribute to the broader literature on how generative AI is reshaping pedagogical and assessment practices, providing empirical evidence for its potential utility in enhancing writing proficiency and engagement in higher education.

The implications of this research extend beyond the immediate context of Hong Kong, offering insights into how generative AI can support educators in managing workloads while delivering high-quality feedback at scale. Machine-generated feedback within analytic systems, such as automated feedback, has been criticized for insufficient attention to how feedback impacts learner affect and motivation [18,19]. Neglecting the motivational and emotional dimensions of feedback increases the likelihood of it being rejected by learners, demotivating them, or reinforcing harmful beliefs [19–24]. This issue is particularly pronounced among struggling learners, for whom feedback can prompt upward comparisons or act as a self-fulfilling prophecy [22,24,25]. By examining the dual dimensions of student outcomes and experiences, this study provides a comprehensive evaluation of generative AI's role in fostering language development and engagement. It highlights not only the promise of this technology but also its limitations, helping to inform best practices for its integration into educational settings.

Background

AI refers to computerised machines and systems that can perform tasks that might otherwise require human intelligence [26]. Recent advancements in technology mean that AI is now more adept at comprehending human language, recognising patterns, and arriving at decisions based on guidance [27]. Its potential application across a number of sectors, including education, has been emphasised by prior research [28], rendering it one of the chief areas of investment into educational technology [29]. Artificial Intelligence in Education (AIEd) is thus a burgeoning sector within educational technology, offering potential benefits for large-scale teaching environments and providing real-time, personalized feedback to students [30].

Despite AI's integration into applications over the past 30 years, ongoing research is essential to support large-scale teaching and intelligent assistance [31]. Natural Language Processing (NLP), a subset of AI, has seen significant advancements in text processing, particularly with the development of transformer-based models like those used in self-attention mechanisms for NLP. Recent advancements in the subset of AI known as Generative AI (GenAI) have rapidly improved the capacity of the technology to both analyse and create content, opening the way to new potential applications for AIEd [32]. Such applications typically rely on larger language models (LLMs), which are trained on large amounts of textual data to generate human-like language, as well as performing tasks such as translation, summarisation, and answering questions [33].

Among LLMs, ChatGPT has emerged as one of the more sophisticated LLMs due to its massive training dataset and advanced architecture, rendering it capable of generating highly versatile, contextualised responses suitable for diverse applications [34]. The advent of powerful LLMs such as ChatGPT suggests a promising future for AIEd [35]. The technical capabilities of automatic assessment systems have improved, and numerous studies highlight the potential of AIEd. Major advancements in AIEd can be categorized into four key areas [31]: decision-making tools, intelligent tutoring systems, adaptive systems, and assessment and evaluation tools.

Decision-making tools aid in profiling and predicting admissions decisions, course scheduling, drop-out and retention rates, student modelling, and academic performance [36–38]. Intelligent tutoring systems are designed to teach course content, interact with students, curate learning materials, facilitate collaboration, and support teachers [39,40]. Adaptive systems offer scaffolding and content personalization, help teachers understand student learning, use academic data to monitor and guide students, and represent knowledge through concept maps [41, 42]. Assessment and evaluation tools are used for automated grading, providing feedback, evaluating student understanding and engagement, ensuring academic integrity, and assessing teaching effectiveness [43, 44]. These areas mark significant progress in AIEd.

This research is premised on the potential utility of feedback in improving student outcomes in written work. Extensive research supports the role of feedback in contributing to improved learning for students [45], and studies have also demonstrated that feedback that informs the revision process can improve grades on written work [46]. However, providing feedback is a time-consuming process for teachers and marking is often cited as a source of teacher workload and stress [47], with low-quality feedback also being a common complaint among university students [48]. This implies the need to improve feedback quality whilst reducing the workload upon teachers. In this regard, automated feedback presents a promising avenue to accomplishing this dual objective [30].

In terms of the potential of automated feedback programmes to increase feedback consistency and reduce teachers' workload, some studies have already investigated the potential of automated writing evaluation (AWE) to reduce marking demands upon teachers [49]. Furthermore, a growing body of research investigates the potential for automated feedback to be applied by computer programmes [50].

However, previous attempts at developing such systems have often focused on task-specific programmes that are naturally limited in terms of their application for teachers on courses where teachers may pose a number of tasks (e.g., offering a choice of essay questions) [51], or where assessment is based on broad criteria (e.g., opinion or reflection based writing tasks) [52]. This is because such technologies have heretofore been limited in terms of making specific assessments towards specific tasks, whereas tasks such as grading open-ended essays requires AI to undertake more complex analysis and assessment [53]. The apparent capacity of GenAI to accomplish this suggests its potential utility towards providing automated feedback on diverse academic work.

Rationale

The rapid improvement in AI technologies across the early 2020s have signalled the prospect for utilising generative AI based on LLMs to both assess written work and to provide written feedback aimed at improving students' writing. As the literature review below reveals, research into the use of generative AI to provide feedback on students' work is growing [54,55], though there has been scant research into its application in the context of university English-language education in Hong Kong. The diverse nature of assessment across various educational systems and cultures means that there are often issues with the generalisability or transferability of findings arrived at in only one context [56]. This means that there is currently insufficient empirical evidence supporting the claim that AI-generated feedback can have a positive effect on student revision to academic work in the context of language education in Hong Kong [52].

As the review of literature below reveals, this research is practically unique within this context, implying its potential to make a major contribution towards the empirical support for a growing area of AIEd in Hong Kong. Through completing research in this area, the utility of generative-LLM AI is evaluated with respect to its applicability in contemporary university-level language education in Hong Kong, taking into account not only its utility towards guiding student revisions, but also with respect to how students respond to GenAI feedback in terms of engagement and motivation. The ramification of this with respect to the extant body of research and the state of AIEd in Hong Kong are reviewed in light of this research in the Discussion section below.

Aims

This study aims at closing a research gap identified with respect to language education in Hong Kong. It is hoped that it is able to (1) assess the potential utility of AI in providing feedback on the written work of university-level language students within this context and (2) contribute to the broader literature on how AI is reshaping pedagogical and assessment practices. It thus aims at making a contribution both to language education within universities in Hong Kong as well as carrying the discussion forward regarding the empirical basis supporting the use of AI in providing feedback on written work more broadly.

Research questions

Meeting these aims requires designing a study that is suitably tailored to closing the gap in knowledge. In this vein, the study's research questions are as follows:

To what extent can LLM-based generative AI provide feedback on written products that improves students' quality of work?
What are the experiences of students when receiving feedback from LLM-based generative AI, particularly in terms of their motivational, emotional, and attitudinal states?

This study responds to these research questions using a mixed-method design, utilising both quantitative and qualitative methods of data collection and analysis. The details of how this is designed and the reasoning behind the specific design of the experiment and research instruments are given in the methodology section below.

Literature review

The extant body of literature on using LLM-based generative AI for feedback suggests there is potential utility for the application of AI to this end. Until relatively recently, studies largely focused on AWEs and their capacity to evaluate student work, though it has also been noted that its capacity for providing individualised feedback was limited [57]. The research findings regarding the effectiveness of providing feedback through LLM-based generative AI in improving student outcomes on revised work were mixed, with limitations observed in the feedback's applicability and specificity to the tasks that the AWE system is designed to evaluate [50].

By way of comparison, LLMS such as GPT have since emerged as a means to provide more tailored feedback to writing products [58]. LLMs are trained on significant amounts of textual data, allowing them to generate natural language that mimics human feedback [59]. They are also capable of providing feedback on different types of work based on task inputs, learning objectives, and scoring systems, requiring relatively less coding time compared to AWEs [60]. There is therefore significant potential with respect to the application of LLMs such as GPT in providing automated feedback [61].

However, as Chang et al. [62] note, there is a lack of studies providing empirical support for the efficacy of the feedback generated by LLMs. Some have expressed concerns that AI-generated feedback might not be accurate given that generative AI such as GPT often makes factual errors when completing generative tasks [63]. Likewise, some have observed that AI feedback in the hands of non-experts may not be as effective as feedback under research conditions given that student prompts might not be sufficiently detailed [64]. On the other hand, LLMs typically perform better at creative tasks, which may include providing feedback [65]. Chang et al. [62] also note that LLMs are capable of providing feedback without the use of reference texts and that they exhibit more potential for feedback than AWE.

This potential for LLMs to provide useful feedback is supported by some empirical studies. In comparing LLM and instructor feedback on written reports produced by university students, one study found that the AI-generated feedback was both coherent and broadly cohered with instructor feedback in terms of positive or negative assessments of the work [66]. Other studies that use student or instructor evaluation of LLM-generated feedback report positive assessments of the technology and its utility on behalf of human participants [67,68]. This may be balanced against some studies on the perspectives of English language teachers who express concerns about linguistic fidelity, overreliance on AI, and the suppression of student creativity [69]. However, there are limited empirical studies into measuring the impact of generative AI on student outcomes.

The few studies that have been carried out report promising results with respect to the utility of generative AI to provide helpful feedback on student work. One study on GPT feedback found that students who used AI to research their work (including using it for feedback) demonstrated better critical, reflective, and creative thinking skills than students who used traditional means of research and feedback [70]. In a study by Meyer et al. [71], 459 upper secondary EFL students were divided into two groups: one group received LLM-generated feedback and the other did not. The findings revealed that the written work evaluated by the Automated Writing Evaluation (AWE) system showed greater improvement in the group that received feedback and revised their work, compared to the group that received non-AI feedback, suggesting the possibility of producing similar findings among ESL students at the university level. Studies have also highlighted the potential of generative AI to transform essay writing and revision processes in higher education, demonstrating significant improvements in essay quality, increased student engagement, enhanced motivation, and mixed emotional responses to AI-generated feedback [72–74]. These insights suggest that generative AI holds considerable promise in supporting student learning and writing development across diverse educational contexts.

Beyond direct learning outcomes in terms of the scoring of written work, there are also other areas where AI feedback may be compared against instructor feedback. For instance, studies have shown that students' beliefs about the value of completing certain English as a Second Language (ESL) writing tasks is linked to their motivation to complete such tasks [75]. Motivation is likewise related to positive student emotions, which have been found to be vital to the process of writing [76], and can be fostered by instructor feedback [77]. In this regard, students need to perceive feedback as effective in order for it to have a positive effect on emotions, motivations and engagement with specific tasks [78].

Fortunately, some preliminary evidence suggests that LLM-based feedback can indeed have a positive effect on student emotions. For instance, one study by Li and Xing [79] demonstrated that LLMs could provide effective emotional support for students. Another study found that interaction with generative AI elicited positive perceptions and high levels of engagement [80], though this was carried out with younger students. One study on the use of GPT with EFL learners found that its reframing of tasks could help foster greater cultural awareness among students, with positive responses from participating students [81]. Importantly, a study by Al Shloul et al. [82], which investigated the potential for GPT to improve student performance through feedback, found that most students saw its feedback as valuable and found interaction engaging.

However, what is less known based on the above studies is whether LLMs can provide feedback in a way that is perceived as effective by students, foster positive emotions, and motivate students to engage in work. This is particularly important given that feedback cycles, revision and submission can prove emotionally draining and demotivating for some students [83]. The study carried out by Meyer et al. [71] found moderate increases in task motivation and positive emotions, indicating the potential for LLM feedback to have beneficial emotional responses for participants. However, there is arguably also a need for qualitative research into these relationships in order to understand what aspects of AI-generated feedback students respond to in a positive (or negative) manner.

This literature review has highlighted several gaps in the literature. First, there are few studies that attempt to demonstrate the efficacy of LLM-based generative AI on student outcomes, though those that have been carried out report positive correlations between the technology and student learning outcomes following revision of work. Importantly, there is a gap in current knowledge as to whether AI-generated feedback can have a positive effect on student revisions to academic work in the context of education in Hong Kong. This highlights the need for more work in this area to establish connections that focus on specific educational contexts and areas of learning. Additionally, early indications that LLM feedback might be used to bolster student emotions and motivation require more qualitative research in order to better understand the mechanisms behind these relationships. These gaps inform the design of this research, as set out below.

Methodology

This study is a true experimental design that employs a randomized controlled structure to evaluate the impact of AI-generated feedback on essay revisions and student engagement. A total of 1102 first-year undergraduate students enrolled in a foundational English course at a higher education institution in Hong Kong participated in the study. Students were randomly assigned to one of two groups: the experimental group ($n = 411$), which received AI-generated feedback provided by ChatGPT, or the control group ($n = 691$), which received general feedback written by an instructor. Randomization was employed to ensure that any differences observed between the two groups could be attributed to the intervention, controlling for potential confounding variables such as baseline writing ability, motivation, and language proficiency. The inclusion of a control group allowed for a direct comparison of outcomes, adhering to the principles of a true experimental design.

The study was conducted over two consecutive semesters, with the same experimental procedure applied to two separate cohorts of students. In each semester, students completed an argumentative essay as part of their coursework. After submitting their initial drafts, students in both groups received feedback specific to their assigned condition (AI-generated or instructor-generated), which they were required to use to revise and resubmit their essays. The independent variable in this study was the type of feedback provided, while the dependent variables included essay scores, motivation, engagement, and emotional responses to the feedback. By carefully manipulating the independent variable and measuring the dependent variables, the study sought to rigorously evaluate the effectiveness of AI-generated feedback compared to traditional instructor feedback.

The rubric used to assess students' essays was specifically developed to align with the course's learning objectives and the study's focus on academic writing. The rubric evaluated three key areas: language use and mechanics, organization and structure, and content and ideas. These categories reflected the skills most likely to be influenced by feedback and were chosen to ensure consistency between the course's instructional goals and the evaluation criteria. The development of the rubric involved adapting established academic writing rubrics widely used in similar contexts, with minor modifications to suit the specific aims of the study. A panel of three experienced English lecturers reviewed the rubric to ensure clarity and relevance. To further validate the rubric, a pilot test was conducted with a small group of students prior to the main study, and inter-rater reliability was assessed using the Intraclass Correlation Coefficient (ICC), resulting in a high coefficient (ICC $= 0.831$). This process ensured that the rubric provided a reliable and objective measure of student writing performance.

To capture students' experiences with the feedback and their emotional and motivational responses, a post-task questionnaire was developed. This questionnaire consisted of Likert-scale items, rated on a scale from 1 to 10, designed to measure key constructs such as positivity, satisfaction, engagement, and confidence during the revision process. The questionnaire was chosen because it allowed for systematic and quantitative comparisons between the experimental and control groups, providing a clear picture of how students perceived and interacted with the feedback. Its design was informed by existing research on feedback and student engagement, drawing on validated tools used in previous studies [77,78]. The questionnaire was piloted with 50 students to ensure clarity and relevance of the items, and based on feedback from the pilot, minor adjustments were made to refine the wording of specific questions. The final version of the questionnaire demonstrated high internal consistency, confirming its reliability for capturing students' emotional and motivational responses.

In addition to the questionnaire, semi-structured interviews were conducted with a subset of 18 students from the experimental group across the two semesters. These interviews aimed to provide rich qualitative data on students' perceptions of the feedback and its impact on their writing process and motivation. The semi-structured format was selected because it offered the flexibility to explore unexpected or nuanced responses while maintaining a consistent framework across participants. The development of the interview schedule was guided by the study's research questions and insights from the broader literature on feedback in education [71,83]. The interview schedule was reviewed by a team of three researchers to ensure its relevance and clarity, and it was piloted with five students to refine the wording and structure of the questions.

Sample

The study collects data from 1102 students enrolled in the first year of an English-language course at a higher education institute in Hong Kong. This course constituted a General English course and was compulsory for all students attending the institution. The course is designed to improve general English ability and all students must also take a course on English for Specific Purposes (ESP) tailored to their degree subject. The course was also designed to assist with International English Language Testing System (IELTS) accreditation, one of the leading standardised tests for assessing English language proficiency for non-native speakers. All students must take a foundational writing skills course, designed to improve students' use of academic English.

The students recruited into this study were studying a variety of disciplines across the Humanities, Social Sciences, and STEM (Science, Technology, Engineering, and Mathematics).. Given the diverse nature of the subjects studied by students enrolled on this course, the course itself is designed to be flexible, with assessments allowing students to choose from a number of questions potentially related to their specific subject. This allows students to incorporate subject-specific vocabulary into their assessed written work. However, some assessment questions are broad and opinion-based, testing students' capacity to engage in written debate on a given topic derived from general knowledge or current affairs. This renders the course suitable for testing the suitability of AI towards offering feedback to students that is not limited to the context of one disciplinary area.

Students were approached to participate through the university ahead of the experiment and were provided with research information as well as consent forms. Emails were sent out to all students enrolled in the course informing them of the intent to conduct an experiment during one of their scheduled tutorials and information and consent forms were again provided in the lesson itself prior to the task. All students participating in the course completed the task, though only those who opted to participate had their data stored for inclusion in the task. The task was carried out individually in tutorial groups of between 20 and 40 students, all of which were overseen by the lecturer and the researcher.

The students participating in the task were all Hong Kong citizens for whom English was a second language, while students who did not meet these criteria were filtered out. Of the samples, 51 % were female and 49 % male, with sorting into the experiment and control groups being as representative as possible of this ratio. In total, 411 students were within the feedback group and 691 students were in the control group, with this ratio selected to reduce resource demands upon the experiment group. Of the students that progressed to the interview stage of the study, 18 students participated, selected randomly from a male and female group of applicants to ensure gender balance in light with the class demographics (see Table 1). These students were scheduled to participate in one-to-one interviews with the researcher within 48 h following their completion of the task.

Table 1
Demographic information for interview participants.

Name	Sex	Age	Year	Field	L1
A	F	18	1	Design	Cant.
B	F	19	1	Design	Cant.
C	F	20	1	Educ.	Cant.
D	F	18	1	Human.	Cant.
E	F	18	1	Edu.	Cant.
F	F	18	1	Science	Cant.
G	F	19	1	Human.	Cant.
H	F	22	1	Design	Cant.
I	F	18	1	Business	Cant.
J	M	18	1	Eng.	Cant.
K	M	19	2	Design	Cant..
L	M	18	1	Business	Mand.
M	M	20	1	Business	Cant.
N	M	18	1	Science	Cant.
O	M	18	1	Science	Cant..
P	M	19	1	Maths	Cant.
Q	M	18	1	Maths	Mand.
R	M	20	1	Human.	Cant.

Experimental design

A number of considerations went in to the specific design of the study. For one, an experimental design was selected that could compare two groups of students, one of whom had received the GenAI intervention and one that had not. The experimental design was selected to help better establish causation in terms of identifying receipt of Gen AI feedback as an independent variable. Similar designs have been used in studies undertaken outside of this context and the study's experimental design was informed by their methods [84]. The large group size and comparable proficiency of students within it made the group suitable for undertaking statistical comparisons of an experimental and control group, whilst also allowing for some replicability in the study design, thereby enhancing the validity and rigour of the research [85].

In addition to this, a qualitative component to the design was added in order to explain the findings. Whilst quantitative results might establish correlations, qualitative analysis can expand upon the ways in which students found GenAI to be useful or not useful [86]. It also allows for the more nuanced exploration of student experiences and perceptions as its data does not need to be reduced down for quantitative analysis. Taking these results together, the qualitative aspect to the study can help to triangulate the findings of the quantitative analysis, thereby strengthening the study's perceived validity should the results align.

Procedures

The experiment was designed to be undertaken during the normal lesson times of students, forming a task administered during a two-hour class held in a computer laboratory on campus. The researcher thus completed multiple sessions so that participants could complete the task during their normal class hours. Test conditions were upheld, meaning that interaction and conferral was prohibited, as was use of AI resources to inform essay writing. Experimental designs in education carried out under test conditions are subject to concerns such as class size and the researcher's capacity to efficiently supervise for cheating or plagiarism [87]. The task took place within classes with an average of 30 students, with a researcher placed in the room alongside the class' usual teacher to supervise the activity and prevent plagiarism and use of AI to complete the task.

The question posed to students was designed in line with the writing tasks provided to students as part of their course. Though these tasks were often subject-specific, this question was more general and was developed in communication with course teachers to be reasonably achievable according to the students' knowledge and level of English proficiency. Previous studies comparing teacher and AI-generated feedback have favoured argumentative essays due to their relatability and lack of dependence upon knowledge of subject-specific terminology [88]. Students were thus asked to complete the following writing task under test conditions:

> Do you agree or disagree with the following statement? Children under five ought to be prohibited from using tablet computers or smartphones. Use specific reasons and examples to support your answer.

Instructions on how to complete the task were given by the researcher at the outset of the study and were repeated above the essay prompt in emails received by the participants.

Students were asked to write an argumentative essay answering the question and defending their reasoning, without recommendation as to word length. A short list of learning objectives and a summary of the rubric was provided to them alongside the notice that students would not be marked on nor penalised for referencing. The learning objectives and rubrics were adapted from those in use on the course itself, thereby accurately reflecting the level of knowledge and skills of the cohort. Students were given 30 min to complete this task and then emailed their responses to the researcher, who would then provide them with their feedback via email whilst they completed another unrelated task as part of their class. Essays received one minute after this cut-off time were not included in the data submitted for analysis.

Those in the feedback group had their work submitted by the researcher to ChatGPT 3.5. ChatGPT was selected for this task because of its capacity to offer personalised feedback without instructor input, as well as its capacity to offer feedback in line with guidance [89]. Submission of student work was thus preceded by a prompt setting out the task instructions and learning objectives and requesting no >500 words of feedback:

A number of students of English studying at the undergraduate level at a university in Hong Kong have been tasked to write an essay on the following question:

> 'Do you agree or disagree with the following statement? Children under five ought to be prohibited from using tablet computers or smartphones. Use specific reasons and examples to support your answer.'

They have been tasked with meeting the following four learning objectives through completion of the assignment:

> '1. Engage critically with the question, meeting the criteria set out in the instructions.
> 2. Argue convincingly towards a clear thesis, drawing on existing research or evidence where possible.
> 3. Employ academic vocabulary and formal English in your answer.
> 4. Ensure accurate spelling and grammatical coherence throughout.'

Based on these learning objectives, please provide no >500 words of feedback designed to help the student improve their essay and raise their grade, giving both general and specific guidance as to areas and means for potential improvement. The essay in question follows below.

At the time where the task recommenced, all students received an email asking them to revise and improve their papers, with students in the feedback group receiving their LLM-generated feedback, and those in the control group receiving the same generalised guidance produced by a teacher beforehand. The students were given 5 min to prepare (and to read their feedback) and then a further 20 min to make revisions, before resubmitting their work to the researcher. Examples of the type of feedback received and its effect on revision may be found in Appendix 1.

Both the original and revised papers were also marked by instructors on the course, who followed a rubric designed for completion of this task (Appendix 2). Manual scoring of work on behalf of experienced instructors was undertaken due to the limited accuracy of LLMs and specifically GPT in scoring written student work [90,91]. Papers and their revised papers were marked by separate instructors and all papers were double-marked with an average of the two marks constituting their final score. In order to ensure the accuracy of the markers, inter-rater reliability was tested through establishing the intraclass correlation coefficient (ICC). This was undertaken to ensure that the scores given by the markers were broadly aligned [92]. Across the markers, an ICC of 0.831 was established, falling within the 95 % confidence intervals of 0.798 and 0.849, and establishing the consistency of the scores awarded.

Quantitative methods

At the end of the experiment, all participating students were asked to fill in a short questionnaire about their experience, completed in the same class immediately after the task. The questionnaire – provided in English at a level suitable for the students' proficiency – was developed by the researcher and focused largely on their emotions and experiences with relation to the process of making revisions, asking them to describe how positive their emotions were during revision, how motivated they were to complete revisions, and how engaged they were with the process of revision (see Appendix 3). Scalar responses were collected that could then be compared across the control groups and measured against their scores from the writing task.

Analysis of the questionnaires took place in IBM's Statistical Package for the Social Sciences (SPSS) 29.0. This allowed for variables to be defined (e.g., gender, scalar variables, etc. and then cases created from data entered into the programme [93]. Tests such as Pearson's product-moment correlation coefficient were used to identify numerical relations between sets of data, whilst the student's *t*-test was used to compare two or more groups' scores across a numerical variable [94]. Point biserial correlation was selected on the one hand because it is suitable for comparing binary data, whereas Pearson's correlation may be useful for exploring additional linear relationships pertaining to revisions [95]. An alpha of 0.05 was used across the tests, whilst the tests themselves were applied to data pertaining to test scores, questionnaire results, etc.

Qualitative methods

Following each experiment session, an interview with a participating student from the feedback group was arranged to discuss their experience of LLM-generated feedback. In total, 18 of these interviews were successfully completed, carried out within 48 h of task completion and lasting around an hour each. Interviews were selected because of their capacity to generate substantial information about individual perspectives as compared with questionnaires [96]. Interviews were carried out by the researcher, who utilised a semi-structured approach to questioning, suitable for not only following the questions but also allowing the researcher to prompt the students for more details about areas of interest [97]. The interviewer conducted these interviews using a protocol (Appendix 4), in line with the standards for semi-structured interviewing [98]. The interviews were recorded on the researcher's tablet computer using digital audio recording software and then transcribed automatically using digital transcription software, before being manually corrected for any transcription errors.

The interview data was then subjected to thematic analysis. Thematic analysis serves as a means for identifying the themes raised by interviewees throughout the research process [99]. It focuses on identifying, analysing and reporting patterns across data, describing and interpreting the themes prevalent across a dataset [100]. As stated above, qualitative analysis was selected in order to help triangulate the study's findings and aid its explanatory analysis, with thematic analysis presenting a useful means for extracting relevant data in this regard [86]. An approach to coding the data is required in order to complete this process [101], with the use of AI itself becoming more common in methods of thematic coding and analysis [102].

In this case, *Leximancer* was selected as a means for conducting thematic analysis of the interview data. *Leximancer* uses algorithms to extract semantic and relational data from the dataset and aggregate them into themes [103]. These are represented in visual charts of ranked and co-occurring concepts [103]. The unsupervised approach to coding and analysis was used in order to allow for the researcher to follow an inductive approach to analysing interviewee responses to interview questions. In this way, *Leximancer* effectively completed extracted themes from the interview texts, which the researcher thereafter has discussed with relation to specific excerpts picked up by *Leximancer*.

Results

This section presents the findings of the mixed-methods approach taken to study within this paper. It presents first the results of the quantitative analysis of data and then the findings of the qualitative analysis of interviews. These findings are discussed in more depth in relation to the study's aims and research questions in the discussion section that follows.

Quantitative analysis

Examining the data for the test scores and questionnaires, it is apparent there are a number of differences between the feedback and control groups. For one, Table 2 illustrates that the revised scores for the feedback group are considerably higher than that of the control group, with the group receiving AI feedback scoring some 3.342 marks higher on this task compared to the earlier task, for which scores were far more comparable. In addition to this, the feedback group reported higher scores on emotion, motivation and engagement, with the difference for self-reported motivation being considerably higher in the feedback group.

These *t*-tests were subjected to a *t*-test in order to establish whether these observed changes in score were statistically significant. This was applied both to the improvement between the two test scores, as well as to the group means for emotion, motivation and engagement scores (Table 3). We conducted *t*-tests to compare the essay scores and self-reported measures between the experimental and control groups. Before conducting the *t*-tests, we assessed the assumptions of normality and equal variances. The normality of the data was tested using the Shapiro-Wilk test, and visual inspections of histograms and Q-Q plots were performed to confirm the results. Homogeneity of variances was assessed using Levene's Test for Equality of Variances. Given the large sample size of this study ($n = 411$ for the experimental group and $n = 691$ for the control group), the *t*-test is robust to minor deviations from normality, as supported by the Central Limit Theorem.

In terms of improvements in test scores, the control group saw an improvement of 4.474 marks (SD = 7.754) and the feedback group a larger improvement of 8.314 (SD = 7.498), resulting in a p value of 0.003106. This p value indicates that there is a high likelihood that the null hypothesis may be rejected. The mean scores for self-reported emotion were more modest, with a mean score of 4.019 in the control group (SD = 2.782) and 4.9755 (SD = 3.119) in the feedback group. The p value here did not fall below the alpha ($p = 0.0785$), indicating that this effect was not statistically significant. However, the values for motivation and engagement were statistically significant and indicated also a larger effect. There was a difference in motivation scores of 1.721 (FG SD = 3.112, CG SD = 2.489, $p = 0.0001$) and a 0.957 difference in engagement score (CG SD = 3.402, CG SD = 2.953, $p = 0.0346$), both in favour of the feedback group. Consequently, it may be said that the improvements to motivation, engagement and revised scores were considerably higher in the feedback group than in the control group and that these effects are statistically significant.

The size of the effects between the groups may be explored beyond the differences between the means above by establishing coefficient scores. A point biserial correlation was calculated for the difference between original and revised paper scores, indicating a weak positive correlation between the group in receipt of AI feedback and increased improvements in test scores (0.208). Comparing the two groups, there was a very weak positive correlation between receiving feedback and reporting positive emotions (0.092) and a slightly stronger but still very week correlation between feedback and engagement (0.155). The correlation between feedback and motivation was higher, however, with a score of 0.31 indicating a weak-to-moderate positive correlation. All observed scores were thus in favour of positive correlations, though statistical significance for these correlations must also be taken into account.

Table 3
T-Test results.

	mean diff.	df	t-value	p-value
Improvements	3.342	230	2.986	0.0031
Emotion	0.630	230	1.564	0.0785
Motivation	1.721	230	4.223	0.0001
Engagement	0.957	230	2.094	0.0346

Attempting to establish the mechanisms at work in the different groups requires understanding how correlated variables such as emotion, motivation and engagement are with the improvements in scores. To this end, Pearson's correlation coefficient was calculated to examine the effect and significance between the above variables. When looking at scores for the feedback and control groups combined, there was an statistically significant moderate effect of emotion upon the test paper score differential ($r = 0.489, p = 1.58678 \times 10–12$). Even stronger were the effects of motivation ($r = 0.901$) and engagement ($r = 0.885$), with both tests reporting high statistical significance ($p = 1.35645 \times 10–68$ and $p = 6.58965 \times 10–71$, respectively). As Fig. 1 shows, the effect of all three experience scores on revised test score improvements increased exponentially, implying that positive affective experiences of feedback and revision became increasingly valuable as they became more enjoyable. It may therefore be theorised that the strong relationships between motivation and engagement and revised scores – coupled with the weak effect of feedback on motivation and engagement – account for the larger relative increase in the scores received for the revised paper among the feedback group.

Additional tests were carried out to confirm these findings, measuring the effect of engagement and motivation on test scores. Simple linear regression tests were carried out to this end for both the feedback and control groups (Tables 4–7). A strong correlation between motivation scores and improvements on test marks were noted, with an improvement of 0.847 and a fit of 70.3 % for the control group, indicating a strong correlation between scores and a high degree of influence of motivation upon these scores. The effect was even higher for the feedback group, however, accounting for some 15 % of the score differentials. Similar but less pronounced effects were noted when calculating the Multiple R for engagement, which increased from 0.879 to 0.911, as well as an R-square increase of 3.3 %. Nevertheless, it seems as though the effect of motivation on test scores was far more pronounced than that of engagement, implying that changes in motivation impact test scores far more than engagement.

Qualitative analysis

Thematic analysis of interviews with eighteen students was carried out through *Leximancer,* using the programme's in-built algorithm to code and organise themes. This method of analysis produces themes based on the frequency, proximity and semantic connections between terms used in the interview transcripts. The four main concepts identified through this process were: effectiveness, experience, impact, and integration. The prevalence of the concepts within the Venn diagram is listed more clearly in Table 8, which signals their frequency across the

Table 2
Average scores for task and questionnaires across and between feedback and control groups.

Group	Task Score	Rev. Score	Difference	Emotion	Motivation	Engagement
Feedback	56.54021	64.85462	8.31441	4.58642	5.07546	4.97554
Control	56.62842	61.51243	4.474227	3.95621	3.35475	4.01891
Difference	−0.08821	3.34219	4.88401	0.63021	1.72071	0.95663

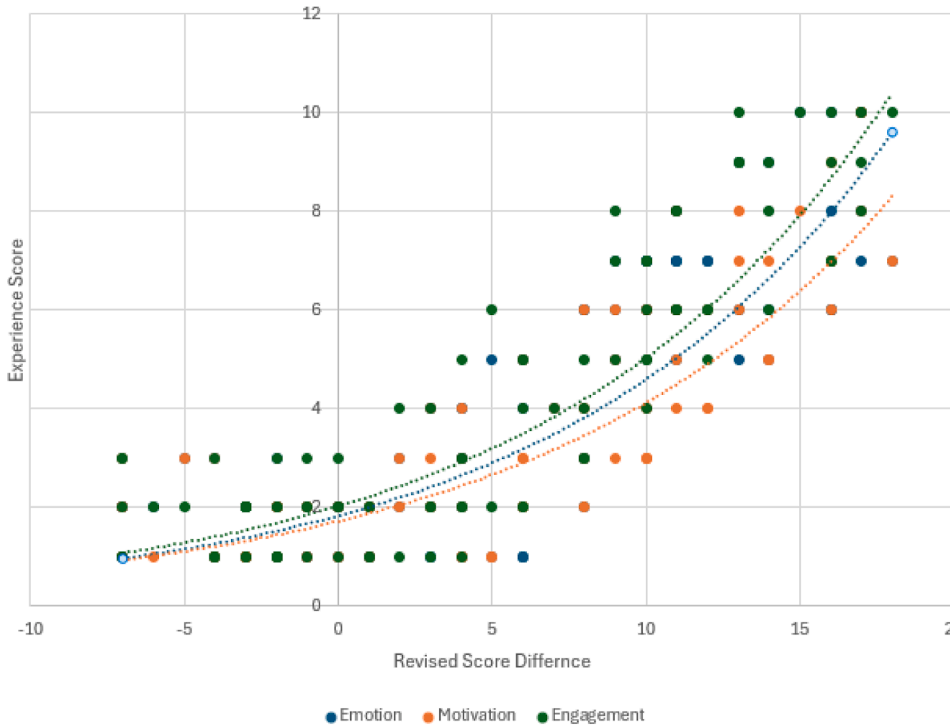

Fig. 1. Self-reported scores for experiences (positive emotions, motivation, and engagement) cross-referenced with associated score improvement on revised task submissions.

Table 4
Regression analysis of motivation effect on score difference for control group.

Multiple R	0.847
R-Squared	70.3 %
Standard Error	1.364
Significance F	3.39×10^{-25}

Table 5
Regression analysis of engagement effect on score difference for control group.

Multiple R	0.879
R-Squared	76.7 %
Standard Error	1.412
Significance F	3.611×10^{-28}

Table 6
Regression analysis of motivation effect on score difference for feedback group.

Multiple R	0.926
R-Squared	85.6 %
Standard Error	1.309
Significance F	3.651×10^{-42}

Table 7
Regression analysis of engagement effect on score difference for feedback group.

Multiple R	0.911
R-Squared	80 %
Standard Error	1.543
Significance F	7.21×10^{-38}

interviews.

These thematic concepts identified in the second table contain themselves other sub-concepts, whose frequency is outlined in Table 9. Here, some related concepts are grouped together – for instance, 'quality', 'improvement' and 'specificity' are grouped under 'effectiveness', whereas 'usability', 'satisfaction' and 'efficiency' are categorised under 'experience'. This indicates the relationship of concepts to each other in the responses, suggesting that the concepts associated with other concepts across the responses. This is developed upon below through looking at concept pathways and the excerpts attached to certain themes identified by the analysis.

The theme of 'effectiveness' was identified with terms such as 'quality', 'improvement', 'specificity', 'engagement', 'accessibility', and 'intuitiveness'. These terms appear to employ effectiveness as a modifier or assessment of aspects to the student experience, such as with respect to improvements upon their revisions and their engagement with work. This is reflected in excerpts from the interviews themselves. For instance, Participant C stated that she felt that the AI feedback was '*effective*… in helping me *improve* the *quality* of my work' (relevant concepts highlighted), with similar statements made with respect to quality of revision throughout. Participant J said of the AI feedback that it 'was generally of high *quality'* and that is reflected an '*improvement* over no help at all', referring to the difficulty of accessing tutor support. The reasons why it was so effective was attributed to its specificity on behalf of the participants, with Participant E stating that 'the feedback was really *specific*', stating that she was 'impressed' with how tailored it was to the work she submitted. Participant H noted also the capacity of the feedback to point out '*specific* errors' with their written English that the participant would not have been able to identify alone.

The theme of 'experience' was associated with terms such as 'usability', 'satisfaction', 'efficiency', 'relevance', 'clarity', 'consistency', 'actionability' and 'impact'. This reflects again potentially positive assessments of the experiences of participants. Participant F stated that she was surprised how precise the feedback was and anticipated that it was be 'easy to *use*' ChatGPT for future attempts at generating feedback to inform her revisions. Participant M reported satisfaction with the

Table 8
Table of major themes with frequency of conceptual frequency.

Theme	Hits
effectiveness	208
experience	150
impact	73
integration	49

Table 9
Ranked concepts derived from Leximancer analysis of interviews with participating students.

	Count	Relevance
effectiveness	193	100%
usability	77	40%
quality	55	28%
satisfaction	40	21%
improvement	36	19%
efficiency	29	15%
specificity	26	13%
relevance	25	13%
engagement	25	13%
clarity	23	12%
confidence	23	12%
accessibility	23	12%
intuitiveness	23	12%
learning	23	12%
consistency	20	10%
actionability	20	10%
impact	19	10%

feedback, stating that it had been a fairly straightforward process to implement: 'The feedback had both general feedback as well as more *targeted* bits where I could actually *apply* it to this or that sentence. Yeah, I was *satisfied* with it'. Participant B also highlighted the clarity of the response offered by the AI feedback: 'Something I liked was that it was very *clear.* […] So I mean that the feedback wasn't general or vague, it told me exactly what I need to do to improve the grade'. The theme of 'actionability' – of receiving guidance that could be put to immediate effect – was highlighted across the responses, with the general experience being one of feedback that was practical rather than abstract. One participant did mention that they felt this was both a strength and failing of the feedback, arguing that the feedback they had received had limited didactic value: 'So I feel like it was good for actually *applying* to my work, but I don't know how much I actually learnt, if you get what I mean'. However, others felt that they had acquired new knowledge from using the feedback. Participant R, for instance, stated that they leant much from the experience of applying the feedback: 'I feel like it's change how I approach looking over my work… Seeing the recommendations under headings has *helped* me think, like, assess your work according to these principles'.

A theme raised by the analysis was that of 'impact', comprising concepts such as 'confidence', 'stress' and 'dependence'. This reflects a mixed assessment as to the perceived impact of the feedback on the student's current or future work. One student stated that they felt more confident completing the revisions with feedback at their disposal: 'I'm a big believer in AI so I guess I was actually quite a lot more *confident* than some of my classmates, as I know some just got a sheet of paper rather than tailored instructions' (Participant L). However, some reported that the feedback had a negative emotional effect, rendering them more anxious or stressed about their work: 'I didn't like getting it. […] Because it's criticising me and that *stressed* me out' (Participant G). It is worth noting, however, that one of the participants who reported feeling stressed stated that she always felt stressed under test conditions, whereas the other reported feeling stress about human feedback also. Participant Q, for instance, felt that a non-human marker reduced the pressure placing upon them when receiving feedback: 'I did feel a bit *stressed* by the feedback, but then I remembered, it's not actually a teacher who can judge you, so that reduced the pressure a little'. A couple of participants reported concerns about potentially becoming dependent upon AI feedback, or alternatively of the institution becoming dependent upon it: 'I would be worried that we sort of devalue getting actual feedback and assume this can do too much' (Participant

K). The concern that having AI capable of checking work immediately might undermine the capacity of students to develop the requisite skills to check their own work was expressed by a minority of those interviewed.

The theme of 'integration' was discussed, incorporating concepts such as 'learning' and 'future'. These concepts suggest that students were prompted to think about how AI could help them with their learning and assessed work in future scenarios. The students gave mixed responses as to how far they felt GenAI could be integrated into their learning with success. One student stated that 'I definitely feel I can *learn* a lot from this', whereas another stated 'I can't see myself *learning* very much, more just correcting things'. A majority of students discussed the prospect of using feedback moving forward, with around half stating that they would again use it to amend their work before submission: 'I'm definitely going to use it in *future* whenever I have an essay because it just structures the feedback in a very useful way' (Participant R). Among those supportive of this proposition, its accessibility and accuracy was highlighted, whilst among those negative about using AI in future, there was scepticism about its accuracy, its contributions to learning, and its efficacy as compared with teacher feedback:

> I would be opposed to receiving it in future, at least through university. I could use it myself if I wanted so having it given to us means nothing. It's not as good as getting a teacher to give you feedback as it's only making corrections, it's not really teaching me anything. And AI can be inaccurate, it even says so on ChatGPT. So I don't really trust it. (Participant Q)

There was thus a mixed response about how GenAI might be integrated into helping with written work moving forwards.

Discussion

The findings identified across the quantitative and qualitative analyses revealed statistically significant correlations between receiving AI-generated feedback and increases in grade between original and revised papers. The effect of this is small but statistically significant, whereas correlations between receiving feedback and motivation and engagement demonstrate a slightly larger effect that is statistically significant. The validity of these findings is supported by the statistical assumptions of the *t*-test, which were carefully tested for normality and equal variances. Minor deviations from normality were mitigated by the large sample size, ensuring robust results. The quantitative analysis further indicates that positive emotions and especially motivation and engagement can have moderate-to-large effects on revision performance, whilst the interviews suggest that effects on motivation through contributing towards confidence are the most pronounced, as the respondents associated it with improvements as compared with receiving generalised feedback. Emotional responses were more mixed, reporting smaller effect sizes that were not always statistically significant, and thus may not necessarily be as correlated with motivations to complete revisions as may otherwise be causally assumed. Although experiences were generally positive, negative assessments of the educational benefits of GenAI on behalf of those interviewed appear to have discouraged a significant proportion from using the technology to this end in future.

The above results demonstrate a number of findings that warrant discussion. Quantitative analysis indicates that the feedback group received higher revised scores and reported higher levels of emotion, motivation and engagement when compared with the control group. T-tests revealed that these differences were statistically significant, with the exception of the differences in emotional responses to feedback. The effect of feedback on positive emotions was also substantially less pronounced according to the point biserial correlation analysis of the effects of feedback on emotion, as well as the Pearson's correlation coefficient examining the relationship between emotional and revision scores generally. This suggests perhaps that positive emotions are not only increased to a lesser degree by receiving AI feedback, but also that their influence on test scores are far less pronounced than motivation and engagement, increases in which must be attributed to factors other than emotion. AI-driven feedback can enhance positive emotions, such as relief and satisfaction, which are associated with better academic performance [106,107]. However, the effect of AI feedback on positive emotions is not as pronounced as its impact on motivation and engagement [108,109]. Positive emotions are increased to a lesser degree by AI feedback compared to other factors like motivation and engagement [108].

The interviews likewise revealed a mixed emotional response to receiving feedback, with participants reporting dissatisfaction with how receiving AI feedback left them feeling disengaged and uninspired. This contradict the findings of similar research, which indicate that LLM-based feedback can a positive effect on student affective experience and emotional well-being that is statistically significant [79]. This may be because feedback by nature is critical, though some did identify a perception of AI feedback as not comparable with human feedback or as lacking in some regard in comparison. Previous studies have also found that feedback cycles can prove emotionally exhausting for students [83], and it may be that this applies more so to LLM-based feedback or was in some way exacerbated by the design of the study.

Interestingly, this did not appear to be reflected in responses regarding motivation, which were broadly positive. An increase in confidence was also linked some the interviewees to improvements made during revision, reflecting perhaps the strong correlations identified by the Pearson's correlation coefficient between motivation and revision scores and engagement and revision scores. Previous research has indicated that motivation is influential with respect to measured outcomes during writing tasks [76], whilst other research has indicated that instructor feedback can successfully enhance student motivation [77]. The above findings seem to suggest that not only can LLM-based feedback enhance motivation significantly, but this also account for a significant proportion of observed differences in test scores. The integration of LLMs in feedback mechanisms has shown potential to improve academic performance by providing clear, structured, and quality feedback, which can help students better understand their strengths and areas for improvement, leading to enhanced learning outcomes [110].

With respect to the effects of AI-generated feedback on these scores, there was also a statistical correlation between receiving AI feedback and the difference between the marks given to the original written product and that to the revised written product. AI-generated feedback has been found to enhance writing outcomes, with students who received AI feedback showing better performance compared to those who received teacher feedback or no feedback at all [111,112]. Using a *t*-test to compare the differences in scores for the control and feedback groups, there was a statistically significant difference between the two groups, with a point bivariate correlation analysis revealing a weak positive effect of receiving feedback on outcomes. This goes some way towards closing the gap observed by Chang et al. [62] regarding the sparse empirical evidence supporting the effect of LLM-based feedback on learning outcomes.

With regard to what may be inferred from the above, it seems to be the case that LLM-based feedback has a positive effect on test scores through providing sufficiently targeted feedback. LLMs help instructors formulate effective feedback by evaluating inputs against quality criteria and generating actionable suggestions, which has been found to improve learning and performance, indicating that LLMs can be a valuable tool in educational settings to provide targeted feedback [110]. That feedback was useful and targeted was barely disputed in the interviews, even among students who reported negative impacts on emotion, implying that its mechanism occurs more through motivation and engagement than through shifts in emotional state. One possible explanation is that the relevant shift in mental states is in terms of attitudinal disposition towards the task, with the interviewees reporting that they felt the feedback gave the task purpose or meaning, as well as citing the usefulness of having specific, actionable feedback at their disposal. This may account for why participants felt the task was useful and helpful though still appeared to compare the task negatively with human feedback. Past research has demonstrated scepticism about GenAI feedback on work, as well as negative assessments when compared with human feedback [113]. It could be that the students felt that GenAI feedback is preferable to *no* feedback, but inferior to expert human feedback.

There is the possibility that these findings are not transferrable into real-world scenarios. Motivation in the context of the task may have been improved by feedback as it provides a purpose for the revision rather than for the utility of its specific guidance. As the tasks were not creditable in terms of contributing to a course of study or qualification, revising the paper may have only appeared to have a point in light of new feedback and 'instructions' from AI. Comparing pre- and post-feedback emotional and attitudinal states is unfortunately not possible as there was no pre-feedback questionnaire, meaning that changes in motivational state before and after the intervention cannot be compared. The nature of the interview analysis also inhibits insights into these relationships given that the output of the algorithm pertains largely to the semantic content of responses rather than the specific views, attitudes and experiences of participants.

Another limitation to the study is with respect to what it says about domain-specific knowledge and written tasks within them. The task assigned to the English-language students was fairly generic and not related clearly to subject-specific skills, which students tasks on the course were typically tailored towards. The respondents to the interviews naturally did not report how far the input from AI improved their writing *skills* and instead focused on its ability to present them with new arguments or content. This perhaps reflects concerns about the limitations of AI in terms of serving well to create content but not as well when it comes to its analytical function [64]. It may be, for example, that the content of the feedback received was not particularly helpful and that its association with relatively improved scores is attributable wholly to its effect on motivation and engagement. In other words, the design of the study does not allow for any evaluation of how accurate, relevant or helpful the actual guidance provided by the AI proved.

A further potential issue with the study design is the limited nature of the task and time for its completion. Restricting the study to a single task per student means that variability in student responses across a range of writing proficiency levels may not have been fully captured. Future studies could utilise a more diverse set of writing tasks to generate a more comprehensive understanding of the effects of GenAI on proficiency. Allowing students only five minutes to consult feedback and prepare their revisions may have also impacted how effectively they were able to make use of their feedback. As the AI feedback was considerably lengthier and more detailed than the instruction provided to the control group, a short period of time for preparation and revisions may have in fact underestimated the effects of GenAI on revisions, or produced more negative emotional responses and reduced engagement due to the imposition of a short timeframe for completion. A longer task in repeat studies may allay such concerns.

Nevertheless, the statistical correlations demonstrate that the group that received feedback from AI did indeed experience improvements in their scores relative to those of the control group. AI feedback in language learning and legal writing contexts improved writing quality and reduced anxiety, outperforming traditional feedback methods [112, 114]. This suggests that whilst there is not yet sufficient information to develop a clear model for pattern of causal influence behind the relationship, there is indeed a positive relationship that warrants further research controlling for variables. This is reflected in the recommendations offered below.

The findings suggest that LLM-based generative AI has the potential to significantly improve the quality of students' written work. The quantitative analysis demonstrates that students who received AI-generated feedback performed better when revising their written papers as compared with the control group, scoring on average 3.113 marks higher. Statistical tests confirm that these improvements were statistically significant and that there was a weak effect ($r = 0.208$) of receiving AI feedback on subsequent performance relative to the control group. Motivation and engagement also had strong positive correlations with score improvements, whilst the feedback group enjoyed a weak improvement in both measures as compared with the control group. Emotional positivity was found to be correlated with scores to a moderate degree but had weaker or insignificant relationships with feedback across other measures.

Students' experiences with LLM-based AI feedback were mixed in terms of the emotional response, with interviews noting a mixed response whilst statistical analysis denied a significant link between feedback and emotional positivity. Interviews indicate that students have diverse emotional responses to AI feedback, with some finding it beneficial and others less so [110,115,116]. Relationships between feedback and motivation and engagement were statistically significant but weak in strength, though both motivation and engagement were strongly correlated with general revision performance across both groups. It is therefore possible that AI feedback functioned through increasing student motivation and encouraging engagement during the task, though it is unclear whether this finding is transferrable to real-world scenarios. Likewise, it is possible that other latent variables play a role in the noted improvements in scores following the intervention. Quantitative data analysis does not support a significant correlation between AI feedback and positive emotional responses, suggesting that while some students may feel positively, this is not a universal experience [117].

Conclusion

This study demonstrates that generative AI feedback can significantly enhance university students' writing proficiency, as shown through improved essay revisions, increased engagement, and heightened motivation during the revision process. The randomized controlled design provides strong evidence that AI-generated feedback can deliver individualized, timely, and actionable input to students, addressing some of the common challenges faced by educators, such as workload and inconsistency of feedback. Meanwhile, educators should help students critically assess AI tools by evaluating feedback quality, understanding limitations, and balancing AI outputs with human judgment [118].

The findings have important implications for higher education, particularly in English-language learning contexts. First, the use of generative AI has the potential to scale personalized feedback in large classrooms, making it a valuable tool for educators seeking to balance quality and efficiency. Second, AI feedback can foster a more engaging learning environment by empowering students to take ownership of their revisions and improving their confidence in writing. These results suggest that integrating generative AI into language education can complement traditional teaching methods and enhance student outcomes. Integrating Generative AI into learning practices enhances the effectiveness of the learning process by enabling teachers to efficiently plan and implement activities aligned with relevant curricula and supported by empirical evidence [119]. Indeed, teachers should take a leading role in designing the curriculum and instructional strategies for AI literacy, ensuring that the content and methods align with practical classroom needs and teaching realities [120].

However, the study is not without limitations. The research was conducted in a specific university context in Hong Kong, which may limit the generalizability of findings to other educational systems or cultural settings. Generative AI demonstrates its potential to effectively complement the sensitivity and contextual awareness of human writers [121–123]. The notion that AI should collaborate with human intellect to enhance, rather than replace, human capabilities—referred to as intelligence augmentation [124,125]—has also been widely supported in educational research on AI-based tools (e.g., [126–128]). Future research could explore the application of generative AI feedback across different disciplines, age groups, and cultural contexts to determine its broader applicability. Additionally, while this study focused on the impact of AI feedback on writing quality and engagement, further research could investigate its long-term effects on students' learning and skill development.

In conclusion, this study highlights the transformative potential of generative AI in higher education. By leveraging its capacity for personalized and timely feedback, educators can improve the quality of student learning while managing workload effectively. Learners, language instructors, researchers, policymakers, and developers should work together to ensure the thoughtful and responsible integration of GenAI tools into language education settings [129]. It is therefore essential for teachers to integrate AI-related skills into their digital teaching competencies, such as understanding how intelligent tutoring systems function to effectively manage automated feedback, enhance learning, and develop suitable support plans for students [64]. As generative AI technologies continue to evolve, their integration into pedagogical practices offers a promising avenue for advancing educational outcomes on a global scale.

Contributions

The development and execution of this study were achieved through the collaborative efforts of all authors, each contributing uniquely. Noble Lo initiated the project, conceptualized the study framework, analyzed the data, and prepared the initial manuscript draft. Alan Wong significantly contributed to the discussion sections, providing critical insights that enhanced the theoretical depth of the study. The findings and theories presented were enriched by joint discussions and brainstorming sessions involving all authors. Sumie Chan was instrumental in data collection, expertly managing data entry into spreadsheets and converting qualitative data into quantitative form, thereby solidifying the empirical foundation of the study. Each author has reviewed and approved the final manuscript, indicating collective agreement on the work conducted and the findings presented.

CRediT authorship contribution statement

Noble Lo: Writing – review & editing, Writing – original draft. **Alan Wong:** Writing – review & editing. **Sumie Chan:** Writing – review & editing, Writing – original draft.

Ethical considerations

Ethical considerations were taken into account when designing this research. For one, the British Educational Research Association's Ethical Guidelines for Educational Research [104] were consulted when designing the study. Following its guidance, all participating students took part in the study voluntarily and were informed fully about their rights to withdraw from the study at any time. Their data was also anonymised at the point of marking and transcription, being attached only to a codename (e.g., Student 1, 2, 3, etc.). Finally, the researcher took into consideration the relative positions of power with respect to their relation to the students throughout the study, reflecting the need for positionality when undertaking primary qualitative research [105].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendices

Appendix 1. Example of student essay, AI-generated feedback, and revised essay

Original Essay

Children in the 21st Century are using tablet computers and smart phones more than ever. A survey of 773 parents in Hong Kong now showed that children are 4 to 6 h every day on electronic devices. Some suggest that this bad for children for reasons such as damage to eyesight, impaired gross motor function, emotional and well-being, and problems with social development. There are now some that say that children should not be allowed to use devices at all under a certain age because of the risks of damage associated with using electronic devices at a young age.

One of the main reasons that this might be a good idea is because children's brains are not formed by the age of 5. Young children's brains are still growing and have not become mature enough to cope with long use of electronic devices. It is argued that children require social action to be develop fully and that using devices all day can cause less time for social action. What's more, smartphones and computers can encourage the wrong type of social action, such as contacting adults on line or seeing adult content. How device use impacts children's health and brains is not well understood scientifically and therefore it is risk to allow children to use so much.

However, there are other arguments to say that using electronic devices is good or okay. For instance, there is more and more work at school from the primary level. Students in the future will probably have to complete work on computers or tablets more than ever, meaning students that don't use these will be left behind in adult world or work. More than ever, student social action takes place on line, meaning that students who don't have phones might be left out socially, hurting their developing. So there are a number of reasons why young children might benefit from using electronic devices.

My opinion is that it is up to parents if young children use devices. Government should guide parents but not tell them how to parent in law. This is part of ensuring freedom in Hong Kong and the world over. Parents should be told about the risks and benefits to young children using phones and tablets through public awareness but should not be made illegal for them to allow their children to use it. This is a fair way of doing things as there is no clear evidence either way as to what young children using electronic devices does.

AI Feedback (ChatGPT 3.5)

This essay addresses an important contemporary issue and presents arguments both for and against the prohibition of young children using electronic devices. While there are some strong points made, improvements are needed in critical engagement, clarity of argument, academic style, and grammatical accuracy in order to raise the overall quality of the work.

1. Engage Critically with the Question

The essay touches on both sides of the debate but lacks depth in critical engagement. You briefly mention potential risks (damage to eyesight, social development) and benefits (preparing children for future work, social inclusion), but these points are underdeveloped.

Suggestions for Improvement:

- Expand on the evidence supporting your claims. For instance, you mention risks such as impaired motor function or social development but do not provide any specific studies or research to support these ideas. Look for scientific studies or expert opinions that discuss these risks in detail.
- Critically evaluate counterarguments rather than just stating them. For example, you could discuss whether preparing children for a digital future outweighs the health risks or explore if limited, supervised use might be a compromise.
- Provide more specific examples, especially for the opposing viewpoint. What kinds of positive educational outcomes might children experience by using tablets or smartphones?

2. Argue Convincingly Towards a Clear Thesis

Your thesis—expressing that parents should decide on the use of devices rather than the government—is introduced at the end but lacks a clear presence throughout the essay. Additionally, the essay feels somewhat unbalanced, with the pros and cons not clearly tied to your overall argument.

Suggestions for Improvement:

- Clarify your thesis in the introduction so that readers know from the outset what your position is.
- Strengthen the coherence of your argument by making sure each paragraph relates to your thesis. For example, the second paragraph could more clearly connect the risks of device use to why you believe the decision should be left to parents.
- Transition between arguments and counterarguments more fluidly. Currently, the sections feel somewhat disconnected, and the essay would benefit from smoother progression.

3. Employ Academic Vocabulary and Formal English

The essay employs some basic academic vocabulary but could benefit from a more formal tone and a wider range of advanced language. Informal phrases like "what's more" and "okay" are more suited to conversational English than academic writing.

Suggestions for Improvement:

- Replace informal phrases with more formal equivalents. For example, instead of "what's more," use "additionally" or "furthermore."
- Use more sophisticated academic language to express complex ideas. For instance, "it is argued" is a good start, but phrases like "it has been suggested by research" or "empirical studies indicate" would strengthen your academic tone.
- Avoid using contractions like "don't" and "it's" in formal writing. Replace them with "do not" and "it is."

4. Ensure Accurate Spelling and Grammatical Coherence

There are several grammatical issues in the essay, particularly with sentence structure and word choice. Errors such as "are 4 to 6 h every day on electronic devices" and "social action to be develop fully" detract from the clarity and flow of your ideas.

Suggestions for Improvement:

- Pay attention to subject-verb agreement and verb forms. For instance, "are 4 to 6 h" should be "spend 4 to 6 h."
- Review your use of articles (e.g., "a risk to allow children to use so much" could be "it is risky to allow children to use them so often").
- Proofread for spelling errors and awkward phrasing. For example, "emotion and well-being" should be "emotional well-being," and "students that don't use these will be left behind in adult world or work" could be rephrased as "students who do not use these devices may struggle in the adult world or workforce."

General Guidance

- Research and Evidence: Incorporate more references to existing studies or expert opinions. This will support your claims and improve the persuasiveness of your argument.
- Structure: Organize your essay more clearly by using topic sentences at the start of each paragraph to signal the argument. Make sure every paragraph ties back to your main thesis.
- Language and Style: Aim to elevate your use of academic vocabulary and avoid informal expressions to maintain a professional tone throughout.

Conclusion

By expanding your arguments, improving grammatical accuracy, and adopting a more academic tone, you can significantly improve the quality of your essay. Clear evidence, stronger engagement with the question, and consistent focus on your thesis will help to create a more convincing and coherent response.

<u>Revised Essay</u>

Children in the 21st Century are using tablet computers and smart phones more than ever. A survey of 773 parents in Hong Kong now showed that children spend 4 to 6 h every day on electronic devices [130]. It has been discussed in the academic discourse that this detrimental for children for reasons such as damage to eyesight, gross motor function, emotional well-being, and social development. There are now some experts that say that children should not be allowed to use devices at all under a certain age because of the risks of damage associated with using electronic devices at a young age. This essay argues that because of the conflicting evidence on the effects of device use in young children, government should guide parents but not ban use of electronic devices by children under 5.

One of the main reasons that banning devices might be a good idea is because empirical studies indicate children's brains are not formed by the age of 5. Young children's brains are still growing and have not become mature enough to cope with long use of electronic devices. Furthermore, it is argued that children require social interaction to develop fully and that using devices all day can cause less time for social interaction. Studies show that children rely on play to develop emotionally and that it is good for their well-being also (Ng, 2023). Additionally, smartphones and computers can encourage the wrong type of social interaction, such as contacting adults online or seeing adult content. There may be physical problems as well, such as affecting eyesight and gross motor function. For example, studies show that during Covid eyesight among young children worsened in Hong Kong. However, there are experts that disagree with these claims and argue that the evidence for long term effects are not established (NCT).

There are other arguments to say that using electronic devices is acceptable or even good for children. For instance, there is more and more work at school from the primary level that is on devices. Students in the future will probably have to complete work on computers or tablets more than ever, meaning students that do not use these devices may struggle in the adult world or workforce. More than ever, student social interaction takes place on line, meaning that students who do not have phones might be left out socially, hurting their development. So there are a number of reasons why young children might benefit from using electronic devices. These reasons need to be critically evaluated so better arrive at a compromise for device use.

Children must be prepared for a digital future as that becomes reality. However, there are also reasons why this is not the most important concern. If children's emotional well-being, social development and physical health is risked, then this digital education can wait until they are more developed. However, the disagreed evidence above indicates that there is not a clear effect on children by early device use. Making this illegal could make parents criminals for letting their children use their phone or watch television on a device. This damages freedom for all and means that government might go too far beyond what evidence suggests is necessary.

In conclusion, government should guide parents but not tell them how to parent in law. Parents should be told about the risks and benefits to young children using phones and tablets through public awareness but it should not be made illegal for them to allow their children to use it. This is a fair way of doing things as there is no clear evidence either way as to what young children using electronic devices will do to them.

Appendix 2. Marking rubric supplied to teachers for marking written task

	Fail (<40 %)	Limited but pass (40–49 %)	Satisfactory (50–59 %)	Comprehensive (60–69 %)	Excellent in most respects (70–79 %)	Outstanding in most respects (80–89 %)	Exceptional in most/ every respect (90 %+)
	0 10 20 32 35 38	42 45 48	52 55 58	62 65 68	72 75 78	84	92 100
Language Use and Mechanics 50 marks	Demonstrates an ineffectual grasp of basic grammar, to the detriment of reader comprehension. Range of vocabulary is limited and inappropriate for context. Shows limited capacity to construct simple sentences. Spelling and punctuation errors may inhibit comprehension.	Demonstrates some grasp of English grammar, though errors may impact understanding in places. Range of vocabulary is basic and may in places not be appropriate for the given context. Shows capacity to construct correct simple sentences. Spelling and punctuation errors may in places distract the reader.	Demonstrates understanding of basic grammatical structures with errors not impacting understanding. Shows an adequate range of vocabulary appropriate for the given context. Can use clear, simple sentences with evidence of more complex structures. Spelling and punctuation errors should not distract the reader.	Demonstrates a basic grasp of grammatical structures with few errors. Employs a good range of vocabulary appropriate to the context. Capacity to construct more complex sentences is evident. Spelling and punctuation errors are few and do not distract the reader.	Demonstrates a strong grasp of grammatical structures with only very infrequent errors. Employs a wide range vocabulary suited to the context. Demonstrates a capacity to construct more complex sentences. Spelling and punctuation are generally accurate and does not inhibit comprehension nor distract the reader.	Demonstrates a very strong grasp of grammatical structures with no errors. Employs a comprehensive vocabulary well suited to the context. Employs complex sentences to effect. Spelling and punctuation are accurate throughout.	Demonstrates an expert grasp of grammatical structures without error. Employs an extensive range of vocabulary that is tailored to the context. Employs complex sentences to effect throughout. Spelling and punctuation are exemplary.
Organisation and Structure 30 marks	Structure is unclear or not present. Paragraphing is not used or is highly erratic. Arguments have little flow or cohesion. Introduction or conclusion are missing.	Structure is apparent though may be erratic or unusual. Paragraphing is present but may be ineffectual. Arguments have some flow or cohesion to them, though frequently digress. Some attempt at introducing ideas at the outset or summarising arguments in conclusion.	Evidence of structure and paragraphing, though ineffectual in places. Arguments have some logical flow and cohesion to them, though may at times digress. Evidence of introducing ideas at outset and summarising arguments in conclusion.	Has clear structure though may not be ideally arranged. Paragraphing is present and broadly consistent throughout. Arguments have a logical flow to them and form part of a cohesive piece. A clear introduction and conclusion are present.	Structure is both present and rational, though could be improved upon. Paragraphing is used consistently to correspond to essay topics. Arguments have a logic flow to them and form part of a cohesive thesis. Includes a comprehensive introduction and conclusion.	Uses a clear, logical structure that sections the content appropriately. Paragraphing is consistent, readable, and corresponds to essay topics. There is a logical flow to argumentation throughout, contributing towards a cohesive thesis. Includes a comprehensive introduction and conclusion.	Is structured according to the conventions of academic work in domain. Paragraphs correspond to a clear topic, are consistently structured, and contribute to the readability of the work. Arguments are arranged to maximal effect, contributing convincingly to the thesis. Introduction and conclusion are wholly comprehensive.

(*continued on next page*)

(*continued*)

	Fail (<40 %)						Limited but pass (40–49 %)			Satisfactory (50–59 %)			Comprehensive (60–69 %)			Excellent in most respects (70–79 %)			Outstanding in most respects (80–89 %)	Exceptional in most/ every respect (90 %+)	
	0	10	20	32	35	38	42	45	48	52	55	58	62	65	68	72	75	78	84	92	100
Content and Ideas 20 marks	Essay is unfocused or demonstrates severe misunderstanding of topic or tendency towards tangentiality. Includes no examples or examples included lack any remote relevance. Ideas are unclear, nonsensical, derivative, or entirely unconvincing.						Some focus on essay prompt, but in places misapprehends or diverges from the topic. Includes at least one example to support ideas though may have limited relevance. Ideas and arguments at times lack clarity, appear to be derivative, or are not wholly convincing.			Some focus on essay prompt, but may not explore the topic in depth or includes tangential points. Includes some examples to evidence points or develop ideas, though may have limited applicability. Ideas and arguments are typically clear and somewhat convincing.			Focused on essay prompt with some depth in discussion and few tangential points. Includes examples to develop ideas that are broadly relevant. Ideas and arguments are clear, somewhat convincing, and some evidence of original thinking.			Focused on essay prompt with in-depth discussion and no tangential discussion. Includes examples to develop ideas that are relevant. Ideas and arguments are clear throughout, are broadly convincing, and suggest some originality of thought.			Focused on essay prompt with no tangential components and demonstrating a very thorough understanding of the topic. Includes examples to develop ideas that are relevant and convincing. Ideas and arguments are clear throughout, convincingly made, and demonstrate originality of thought.	Focused on essay prompt, demonstrating an accomplished understanding of the topic, and containing in-depth analysis and discussion. Includes examples to develop ideas that are relevant, novel, and highly convincing. Ideas and arguments are always clear, wholly convincing, and suggest significant originality of thought.	

Appendix 3. Post-experiment questionnaire delivered to participants

<u>Post-Experimental Questionnaire</u>

Thank you for participating in this research. Please take the time to answer a few questions about your experience. Your responses will help the researchers better understand how students engage with revising their work.

Below is a series of statements about your experience and attitudes towards the revision undertaken and we ask that you indicate the extent to which you agree or disagree with each statement. Please rate each statement on a scale from 1 to 10 (where 1 is strongly disagree and 10 is strongly agree) by circling the appropriate number corresponding with your outlook.

Your responses will be anonymised prior to analysis and will be used solely for the purposes of statistical research, so please answer as honestly as possible.

1. I felt positive about the process of making revisions to my work.									
1	2	3	4	5	6	7	8	9	10
2. I was motivated to complete revisions on my work.									
1	2	3	4	5	6	7	8	9	10
3. I found the process of completing revisions to be satisfying.									
1	2	3	4	5	6	7	8	9	10
4. The process of revising my work was engaging.									
1	2	3	4	5	6	7	8	9	10
5. I felt confident in my ability to make revisions that would improve my work.									
1	2	3	4	5	6	7	8	9	10
6. I found it easy to concentrate when revising my work.									
1	2	3	4	5	6	7	8	9	10
7. I knew which areas of my work would benefit from revision.									
1	2	3	4	5	6	7	8	9	10
8. I felt that the time put into revising my work was worthwhile.									
1	2	3	4	5	6	7	8	9	10
9. The feedback I received helped me feel more motivated to revise my work.									
1	2	3	4	5	6	7	8	9	10
10. I enjoyed the process of revising my work.									
1	2	3	4	5	6	7	8	9	10

Appendix 4. Post-experiment interviewed protocol

Thank you for taking the time with me today.

We're here because you'd indicated previously that you'd be willing to participate in an interview about the writing task we carried out earlier. You of course have the right to withdraw from this study at any time, including after this interview has been carried out. I just wanted to check if you're still happy to proceed? [Proceed if affirmative]

I'm going to ask you some questions about the task, your experience of it, and some questions more generally about technology and its use in education. There are no right or wrong answers, so please be as honest as possible. We've set aside up to an hour for this interview and there are only eight questions, so please take your time and be detailed in your responses. If I need to know a bit more about something, I'll let you know.

1. What were your initial thoughts when you received the AI feedback on your essay?
2. How did the feedback influence your motivation to revise your essay?
3. In what ways did using the feedback affect your engagement with the task?
4. How did you find the AI feedback affected the revisions you made to your work?
5. Can you describe your emotional responses when receiving and applying the feedback?
6. How does the experience of receiving AI feedback compare to the feedback you've received from teachers?
7. How satisfied were you with the process of receiving AI feedback on your work?
8. Do you see yourself using AI tools like ChatGPT to help revise your work in future?

That's all the questions I have for today. Is there anything else you'd like to add that we perhaps didn't cover here?

That's great. Thank you for your time here today. If you've any further questions about the study, your role in it, or if you'd like to withdraw your data, please don't hesitate to contact me through the email provided.

References

[1] Timotheou S, Miliou O, Dimitriadis Y, Sobrino SV, Giannoutsou N, Cachia R, Monés AM, Ioannou A. Impacts of digital technologies on education and factors influencing schools' digital capacity and transformation: a literature review. Educ Inf Technol (Dordr) 2023;28(6):6695–726. https://doi.org/10.1007/s10639-022-11431-8.

[2] Mhlanga D. Digital transformation in education: relevant paradigms and theories of teaching and learning in the industry 4.0. Intelligent systems in digital transformation: theory and applications. Springer International Publishing; 2022. p. 453–70.

[3] Conroy G, Mallapaty S. How China created AI model DeepSeek and shocked the world. Nature 2025. https://doi.org/10.1038/d41586-025-00259-0.

[4] Gibney E. China's cheap, open AI model DeepSeek thrills scientists. Nature 2025. https://doi.org/10.1038/d41586-025-00229-6.

[5] Holmes W, Bialik M, Fadel C. Artificial intelligence in education: promises and implications for teaching and learning. The Center for Curriculum Redesign; 2022.

[6] Farrokhnia M, Banihashem SK, Noroozi O, Wals A. A SWOT analysis of ChatGPT: implications for educational practice and research. Innov Educ Teach Int 2024;61 (3):460–74. https://doi.org/10.1080/14703297.2023.2195846.

[7] Su Y, Lin Y, Lai C. Collaborating with ChatGPT in argumentative writing classrooms. Assess Writ 2023;57:100752. https://doi.org/10.1016/j.asw.2023.100752.

[8] Escalante J, Pack A, Barrett A. AI-generated feedback on writing: insights into efficacy and ENL student preference. Int J Educ Technol High Educ 2023;20(1): 57. https://doi.org/10.1186/s41239-023-00425-2.

[9] Lingard L. Writing with ChatGPT: an illustration of its capacity, limitations & implications for academic writers. Perspect Med Educ 2023;12(1):261–70. https://doi.org/10.5334/pme.1072.

[10] Wang C, Chen X, Yu T, Liu Y, Jing Y. Education reform and change driven by digital technology: a bibliometric study from a global perspective. Human Soc Sci Commun 2024;11(1):256. https://doi.org/10.1057/s41599-024-02717-y.

[11] Narciss S. Feedback strategies for interactive learning tasks. Handbook of research on educational communications and technology. Routledge; 2008. p. 125–43.

[12] Narciss S, Huth K. How to design informative tutoring feedback for multimedia learning. Instructional Design for Multimedia Learning. 2004. p. 181195.

[13] Nelson MM, Schunn CD. The nature of feedback: how different types of peer feedback affect writing performance. Instr Sci 2009;37:375–401.

[14] Graesser AC, D'Mello S. Emotions during the learning of difficult material. Psychology of learning and motivation, 57. Elsevier; 2012. p. 183–225.

[15] Hascher T. Learning and emotion: perspectives for theory and research. Eur Educ Res J 2010;9(1):13–28.

[16] Liu DYT, Bartimote-Aufflick K, Pardo A, Bridgeman AJ. Data-driven personalization of student learning support in higher education. A. Peña-Ayala (ed). Learning analytics: fundaments, applications, and trends, 94. Springer International Publishing; 2017. p. 143–69. https://doi.org/10.1007/978-3-319-52977-6_5.

[17] Merrill Jr K, Kim J, Collins C. AI companions for lonely individuals and the role of social presence. Commun Res Rep 2022;39(2):93–103.

[18] Lim LA, Dawson S, Gašević D, Joksimović S, Pardo A, Fudge A, Gentili S. Students' perceptions of, and emotional responses to, personalised learning analytics-based feedback: an exploratory study of four courses. Assess Eval High Educ 2021;46(3):339–59. https://doi.org/10.1080/02602938.2020.1782831.

[19] Lonn S, Aguilar SJ, Teasley SD. Investigating student motivation in the context of a learning analytics intervention during a summer bridge program. Comput Human Behav 2015;47:90–7. https://doi.org/10.1016/j.chb.2014.07.013.

[20] Gan Z, An Z, Liu F. Teacher feedback practices, student feedback motivation, and feedback behavior: how are they associated with learning outcomes? Front Psychol 2021;12:697045. https://doi.org/10.3389/fpsyg.2021.697045.

[21] Higgins R, Hartley P, Skelton A. Getting the message across: the problem of communicating assessment feedback. Teach High Educ 2001;6(2):269–74. https://doi.org/10.1080/13562510120045230.

[22] Jones KML, McCoy C. Reconsidering data in learning analytics: opportunities for critical research using a documentation studies framework. Learn Media Technol 2019;44(1):52–63. https://doi.org/10.1080/17439884.2018.1556216.

[23] Pitt E, Norton L. Now that's the feedback I want!'Students' reactions to feedback on graded work and what they do with it. Assess Eval High Educ 2017;42(4): 499–516.

[24] Rogers T, Feller A. Discouraged by peer excellence: exposure to exemplary peer performance causes quitting. Psychol Sci 2016;27(3):365–74. https://doi.org/10.1177/0956797615623770.

[25] Roberts LD, Howell JA, Seaman K, Gibson DC. Student attitudes toward learning analytics in higher education: the Fitbit version of the learning world. Front Psychol 2016;7. https://doi.org/10.3389/fpsyg.2016.01959.

[26] Russell S, Norvig P. Artificial intelligence: a modern approach. Hoboken: Pearson; 2021.

[27] Kaplan A, Haenlein M. Siri, Siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Bus Horiz 2019;62:15–25. Volume.

[28] Zhang K, Aslan AB. AI technologies for education: recent research & future directions. Comput Educ Artif Intell 2021. https://doi.org/10.1016/j.caeai.2021.100025.

[29] Sukhanova, K., 2024. *AI in education market statistics – Key trends & figures for 2024.* [Online] Available at: https://techreport.com/statistics/software-web/ai-in-education-market-statistics-2024/[Accessed 31 October 2024].

[30] Gao R, Merzdorf H, Anwar S, Hipwell M, Srinivasa A. Automatic assessment of text-based responses in post-secondary education. Comput Educ Artif Intell 2024; 6. https://doi.org/10.1016/j.caeai.2024.100206.

[31] Zawacki-Richter O. Systematic review of research on artificial intelligence applications in higher education – where are the educators? Int J Educ Technol High Educ 2019:16–39. https://doi.org/10.1186/s41239-019-0171-0.

[32] Bell, E., 2024. *Generative AI: how it works and recent transformative developments.* [Online] Available at: https://www.investopedia.com/generative-ai-7497939 [Accessed 31 October 2024].

[33] Blank I. What are large language models supposed to model? Trends Cogn Sci 2023;27(11):987–9.

[34] Nield, D., 2023. *How ChatGPT and other LLMs work—and where they could go next.* [Online] Available at: https://www.wired.com/story/how-chatgpt-works-large-language-model/ [Accessed 30 October 2024].

[35] Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ 2023;103:102274. https://doi.org/10.1016/j.lindif.2023.102274.

[36] Alvero A, Arthurs N, Antonio AL, Domingue BW, Gebre-Medhin B, Giebel S, et al. AI and holistic review: informing human reading in college admissions. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society; 2020. https://doi.org/10.1145/3375627.3375871.

[37] Chen L, Chen P, Lin Z. Artificial intelligence in education: a review. IEEE Access 2020;8:75264–78. https://doi.org/10.1109/ACCESS.2020.2988510.

[38] Langley P. An integrative framework for artificial intelligence. In: Proceedings of the AAAI conference on artificial intelligence; 2019. https://doi.org/10.1609/aaai.v33i01.33019670.

[39] Feng S, Law N. Mapping artificial intelligence in education research: a network-based keyword analysis. Int J Artif Intell Educ 2021;31:277–303. https://doi.org/10.1007/s40593-021-00244-4.

[40] Hwang G-J, Xie H, Wah BW, Gašević D. Vision, challenges, roles and research issues of artificial intelligence in education, 1. Elsevier; 2020, 100001. https://doi.org/10.1016/j.caeai.2020.100001.

[41] Chen S-M, Bai S-M. Using data mining techniques to automatically construct concept maps for adaptive learning systems. Expert Syst Appl 2010;37(6):4496–503. https://doi.org/10.1016/j.eswa.2009.12.060.

[42] Kabudi T, Pappas I, Olsen DH. AI-enabled adaptive learning systems: a systematic mapping of the literature. Comput Educ Artif Intell 2021;2:100017. https://doi.org/10.1016/j.caeai.2021.100017.

[43] Huang AY, Lu OH, Yang SJ. Effects of artificial intelligence-enabled personalized recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom. Comput Educ 2023;194:104684. https://doi.org/10.1016/j.compedu.2022.104684.

[44] Luckin R. Towards artificial intelligence-based assessment systems. Nat Hum Behav 2017;1(3):0028. https://doi.org/10.1038/s41562-016-0028.

[45] Graham S, Hebert M, Harris K. Formative assessment and writing. Elem Sch J 2015;115(4):523–47. https://doi.org/10.1086/681947.

[46] Gnepp J, Klayman J, Williamson I, Barlas S. The future of feedback: motivating performance improvement through future-focused feedback. PLoS One 2020;15(6). 10.1371%2Fjournal.pone.0234444.

[47] Jerrim J, Allen R. The health of teachers in England over the past 25 years. London: Nuffield Foundation; 2022.

[48] Madigan D, Kim L. Does teacher burnout affect students? A systematic review of its association with academic achievement and student-reported outcomes. Int J Educ Res 2021;105. https://doi.org/10.1016/j.ijer.2020.101714.

[49] Zhang S. Review of automated writing evaluation systems. J China Comput-Assist Lang Learn 2021;1(1):170–6.

[50] Fleckenstein J, Liebenow L, Meyer J. Automated feedback and writing: a multi-level meta-analysis of effects on students' performance. Front Artif Intell 2023;6. https://doi.org/10.3389/frai.2023.1162454.

[51] Botelho A, Baral S, Erickson JA, Benachamardi P, Heffernan NT. Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. J Comput Assist Learn 2023;39(3):823–40. https://doi.org/10.1111/jcal.12793.

[52] Ramesh D, Sanampudi S. An automated essay scoring systems: a systematic literature review. Artif Intell Rev 2022;55:2495–527. https://doi.org/10.1007/s10462-021-10068-2.

[53] Wei P, Wang X, Dong H. The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: a randomised controlled trial. Front Psychol 2023;14:1249991.

[54] Shetty S, Guruvyas K, Patil P, Acharya J. Essay scoring systems using AI and feature extraction: a review. In: Proceedings of the third international conference on communication, computing and electronics systems; 2022. https://doi.org/10.1007/978-981-16-8862-1_4.

[55] Yeung W, Qi C, Xiao J, Wong F. *Evaluating the effectiveness of AI-based grading tools in the summative assessment of higher education.* Seville. In: Proceedings of the 16th annual international conference of education, research and innovation; 2023. https://doi.org/10.21125/iceri.2023.2063.

[56] Hanushek E. Addressing cross-national generalisability in educational impact evaluation. Int J Educ Dev 2021;80. https://doi.org/10.1016/j.ijedudev.2020.102318.

[57] Mertens U, Finn B, Lindner M. Effects of computer-based feedback on lower- and higher-order learning outcomes: a network meta-analysis. J Educ Psychol 2022;114(8):1743–72. https://doi.org/10.1037/edu0000764.

[58] Yang, S., Nachum, O., Du, Y., Wei, J., Abbeel, P., & Schuurmans, D. (2023). Foundation models for decision making: problems, methods, and opportunities. ***arXiv***. 10.48550/arXiv.2303.04129.

[59] Bowman, S. (2023). Eight things to know about large language models. ***arXiv***. 10.48550/arXiv.2304.00612.

[60] Bressane A, Zwirn D, Essiptchouk A, Saraiva A, de Campos Carvalho F, Formiga J, et al. Understanding the role of study strategies and learning disabilities on student academic performance to enhance educational approaches: a proposal using artificial intelligence. Comput Educ Artif Intell 2024;6. https://doi.org/10.1016/j.caeai.2023.100196.

[61] Whalen J, Mouza C. ChatGPT: a revolutionary tool for teaching and learning mathematics. EURASIA J Math Sci Tech editor 2023;19(7). https://doi.org/10.29333/ejmste/13272.

[62] Chang Y, Wang X, Wang J, Wu Y, Zhu K, Chen H, et al. A survey on evolution of large language models. ACM Trans Intell Syst Technol 2024;15(3):39. https://doi.org/10.48550/arXiv.2404.14387.

[63] Lee D, Arnold M, Srivastava A, Plastow K, Strwlan P, Ploeckl F, et al. The impact of generative AI on higher education learning and teaching: a study of educators' perspectives. Comput Educ Artif Intell 2024;6. https://doi.org/10.1016/j.caeai.2024.100221.

[64] Knoth N, Decker M, Laupichler MC, Pinski M, Buchholtz N, Bata K, Schultz B. Developing a holistic AI literacy assessment matrix – bridging generic, domain-specific, and ethical competencies. Comput Educ Open 2024;6:100177. https://doi.org/10.1016/j.caeo.2024.100177.

[65] Chia, Y., Hong, P., Bing, L., & Pira, S. (2023). Instructeval: towards holistic evaluation of instruction-tuned large language models. ***arXiv***. 10.48550/arXiv.2306.04757.

[66] Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y., Gasevic, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. *IEEE Explore*. 10.1109/ICALT58122.2023.00100.

[67] Jacobsen L, Weber K. The promises and pitfalls of ChatGPT as a feedback provider in higher education: an exploratory study of prompt engineering and the quality of AI-driven feedback. OSF Preprints; 2023. https://doi.org/10.31219/osf.io/cr257.

[68] Steiss J, Powell Tate T, GRaham S, Cruz J, Hebert M, Wang J, et al. Comparing the quality of human and ChatGPT feedback on students' writing. Learn Instr 2024;91. https://doi.org/10.1016/j.learninstruc.2024.101894.

[69] Al-Khreseh M. Bridging technology and pedagogy from a global lens: teachers'. Comput Educ Artif Intell 2024;6. https://doi.org/10.1016/j.caeai.2024.100218.

[70] Essel H, Vlachopoulos D, Essuman A, Amankwa J. ChatGPT effects on cognitive skills of undergraduate students: receiving instant responses from AI-based conversational large language models (LLMs). Comput Educ Artif Intell 2024;6. https://doi.org/10.1016/j.caeai.2023.100198.

[71] Meyer J, Jansen T, Schiller R, Liebenow L, Steinbach M, Horbach A, Fleckenstein J. Using LLMs to bring evidence-based feedback into the classroom: aI-generated feedback increases secondary students' text revision, motivation, and positive emotions. Comput Educ Artif Intell 2024;6. https://doi.org/10.1016/j.caeai.2023.100199.

[72] Chan S, Lo N, Wong A. Generative AI and essay writing: impacts of automated feedback on revision performance and engagement. Reflections 2024;31(3):1249–84. https://doi.org/10.61508/refl.v31i3.277514.

[73] Chan S, Lo N, Wong A. Leveraging generative AI for enhancing university-level English writing: comparative insights on automated feedback and student engagement. Cogent Educ 2024;12(1). https://doi.org/10.1080/2331186X.2024.2440182.

[74] Chan S, Lo N, Wong A. Enhancing university level English proficiency with generative AI: empirical insights into automated feedback and learning outcomes. Contemp Educ Technol 2024;16(4):ep541. https://doi.org/10.30935/cedtech/15607.

[75] Eccles J, Wigfield A. From expectancy-value theory to situated expectancy-value theory: a developmental, social cognitive, and sociocultural perspective on motivation. Contemp Educ Psychol 2020;61. https://doi.org/10.1016/j.cedpsych.2020.101859.

[76] Schrader C, Kalyuga S. Linking students' emotions to engagement and writing performance when learning Japanese letters with a pen-based tablet: an investigation based on individual pen pressure parameters. Int J Hum Comput Stud 2020;135. https://doi.org/10.1016/j.ijhcs.2019.102374.

[77] Lipnevich A, Murano D, Krannich M, Goetz T. Should I grade or should I comment: links among feedback, emotions, and performance. Learn Individ Differ 2021;89. https://doi.org/10.1016/j.lindif.2021.102020.

[78] Pandero E, Lipnevich A. A review of feedback models and typologies: towards an integrative model of feedback elements. Educ Res Rev 2022;35(5). https://doi.org/10.1016/j.edurev.2021.100416.

[79] Li C, Xing W. Natural language generation using deep learning to support MOOC learners. Int J Artif Intell Educ 2021;31:186–214. https://doi.org/10.1007/s40593-020-00235-x.

[80] Aslan S, Durham L, Alyuz N, Okur E, Sharma S, Savur C, Nachman L. Immersive multi-modal pedagogical conversational artificial intelligence for early childhood education: an exploratory case study in the wild. Comput Educ Artif Intell 2024;6. https://doi.org/10.1016/j.caeai.2024.100220.

[81] Zheng Y, Stewart N. Improving EFL students' cultural awareness: reframing moral dilemmatic stories with ChatGPT. Comput Educ Artif Intell 2024;6. https://doi.org/10.1016/j.caeai.2024.100223.

[82] Al Shloul T, Mazhar T, Abbas Q, Iqbal M, Ghadi Y, Shahzad T, et al. Role of activity-based learning and ChatGPT on students' performance in education. Comput Educ Artif Intell 2024;6. https://doi.org/10.1016/j.caeai.2024.100219.

[83] McGarrell H, Verbeem J. Motivating revision of drafts through formative feedback. ELT J 2007;61(3):228–36. https://doi.org/10.1093/elt/ccm030.

[84] Wetzler EL, Cassidy KS, Jones MJ, Frazier CR, Korbut NA, Sims CM, Bowen SS, Wood M. Grading the graders: comparing generative AI and human assessment in essay evaluation. Teach Psychol 2024;0(0). https://doi.org/10.1177/00986283241282696.

[85] Andrade C. Sample size and its importance in research. Indian J Psychol Med 2020;42(1):102–3.

[86] Ivankova N, Creswell J, Stick S. Using mixed-methods sequential explanatory design: from theory to practice. Field Mtethods 2006;18(3).

[87] Ross S, Morrison G. Experimental research methods. In: Jonassen D, Driscoll M, editors. Handbook of research on educational communications and technology. London: Routledge; 2004.

[88] Banihashem SK, Kerman NT, Noroozi O, Moon J, Drachsler H. Feedback sources in essay writing: peer-generated or AI-generated feedback? Int J Educ Technol High Educ 2024;21(1):23. https://doi.org/10.1186/s41239-024-00455-4.

[89] Mahapatra S. Impact of ChatGPT on ESL students' academic writing skills: a mixed methods intervention study. Smart Learn Environ 2024;11:9.

[90] Lee G, Latif E, Wu X, Liu N, Zhai X. Applying large language models and chain-of-thought for automatic scoring. Comput Educ Artif Intell 2024;6. https://doi.org/10.1016/j.caeai.2024.100213.

[91] Misiejuk K, Kalissa R, Scianna J. Augmenting assessment with AI coding of online student discourse. Comput Educ Artif Intell 2024;6. https://doi.org/10.1016/j.caeai.2024.100216.

[92] Koch G. Intraclass correlation coefficient. Encyclopedia of statistical sciences, vol. 4. New York: John Wiley; 1982. p. 213–7.

[93] Salcedo J, McCormick K. SPSS statistics. 4th ed. John Wiley; 2020.

[94] McCormick K. SPSS statistics for dummies. John Wiley; 2015.
[95] Kombrot D. Point biserial correlation. London: Wiley StatsRef; 2005.
[96] Peters K, Halcomb E. Interviews in qualitative research. Nurse Res 2015;22(4): 6–7. https://doi.org/10.7748/nr.22.4.6.s2.
[97] Magaldi D, Berler M. Semi-structured interviews. In: Zeigler-Hill V, Shackelford T, editors. Encyclopedia of personality and individual differences. London: Springer; 2020. p. 4825–30.
[98] Kallio H, Pietila A, Johnson M, Docent M. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. J Adv Nurs 2016;72(12):2954–65.
[99] Attride-Stirling J. Thematic networks: an analytical tool for qualitative research. Comm Health Improv 2001;1(3):385–405. https://doi.org/10.1177/146879410100100307.
[100] Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006; 3(2):77–101. https://doi.org/10.1191/1478088706qp063oa.
[101] Guthrie G. Basic research methods: an entry to social science research. Sage; 2010.
[102] Christou P. Thematic analysis through Artificial Intelligence (AI). Qual Rep 2024; 29(2):560–76.
[103] Smith A, Humphreys M. Evaluation of unsupervised semantic mapping of natural. Behav Res Methods 2006;38(2):262–79. https://doi.org/10.3758/BF03192778.
[104] British Educational Research Association. Ethical guideliens for educational research, fourth edition. London: British Educational Research Association; 2018.
[105] Holmes A. Researcher positionality – a consideration of its influence and place in qualitative research – a new researcher guide. Shanlax Int J Educ 2020;8(4):1–10. https://doi.org/10.34293/education.v8i4.3232.
[106] Dehbozorgi N, Kunuku M. Exploring the influence of emotional states in peer interactions on students' academic performance. IEEE Trans Ed 2024;67:405–12. https://doi.org/10.1109/TE.2023.3335171.
[107] Gao S. Can artificial intelligence give a hand to open and distributed learning? A probe into the State of undergraduate students' academic emotions and test anxiety in learning via ChatGPT. Int Rev Res Open Distrib Learn 2024;25(3): 199–218. https://doi.org/10.19173/irrodl.v25i3.7742.
[108] Frey A, Liu T, Fink A, König C. Meta-analysis of the effects of computerized adaptive testing on the motivation and emotion of examinees. Eur J Psychol Assess 2024;40(5):327–443. https://doi.org/10.1027/1015-5759/a000821.
[109] Liao M, Zhu K, Wang G. Can human-machine feedback in a smart learning environment enhance learners' learning performance? A meta-analysis. Front Psychol 2024;14. https://doi.org/10.3389/fpsyg.2023.1288503.
[110] Villagrán I, Hernández R, Schuit G, Neyem A, Fuentes-Cimma J, Miranda C, Hilliger I, Durán V, Escalona G, Varas J. Implementing artificial intelligence in physiotherapy education: a case study on the use of large language models (LLM) to enhance feedback. IEEE Trans Learn Technol 2024;17:2079–90. https://doi.org/10.1109/TLT.2024.3450210.
[111] Li J, Jangamreddy NK, Hisamoto R, Bhansali R, Dyda A, Zaphir L, Glencross M. AI-assisted marking: functionality and limitations of ChatGPT in written assessment evaluation. Aust J Educ Technol 2024;40(4):56–72. https://doi.org/10.14742/ajet.9463.
[112] Wang D. Teacher- versus AI-generated (Poe Application) corrective feedback and language learners' writing anxiety, complexity, fluency, and accuracy. Int Rev Res Open Distrib Learn 2024;25(3):37–56. https://doi.org/10.19173/irrodl.v25i3.7646.
[113] Nazaretsky T, Mejia-Domenzain P, Swamy V, Frej J, Käser T. AI or human? Evaluating student feedback perceptions in higher education. In: *Proceedings of the technology enhanced learning for inclusive and equitable quality education*; 2024. p. 284–98.
[114] Weber F, Wambsganss T, Söllner M. Enhancing legal writing skills: the impact of formative feedback in a hybrid intelligence learning environment. Br J Educ Technol 2024;00:1–28. https://doi.org/10.1111/bjet.13529.
[115] Fong A, Lin J, Ryan L, Chang M. Leveraging the power of generative AI: a case study on feedback analysis of student evaluation in an undergraduate physiology practical course. Physiology 2024;39(S1):2081. https://doi.org/10.1152/physiol.2024.39.S1.2081.
[116] Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, Mahling M. A language model–powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ 2024;10. https://doi.org/10.2196/59213.
[117] Horvers A, Molenaar I, Van Der West H, Bosse T, Lazonder A. Multimodal measurements enhance insights into emotional responses to immediate feedback. Front Psychol 2024;14. https://doi.org/10.3389/fpsyg.2023.1294386.
[118] Tzirides AO, Zapata G, Kastania NP, Saini AK, Castro V, Ismael SA, You Y-L, Santos TAD, Searsmith D, O'Brien C, Cope B, Kalantzis M. Combining human and artificial intelligence for enhanced AI literacy in higher education. Comput Educ Open 2024;6:100184. https://doi.org/10.1016/j.caeo.2024.100184.
[119] Pesovski I, Santos R, Henriques R, Trajkovik V. Generative AI for customizable learning experiences. Sustainability 2024;16(7):3034. https://doi.org/10.3390/su16073034.
[120] Sperling K, Stenberg C-J, McGrath C, Åkerfeldt A, Heintz F, Stenliden L. In search of artificial intelligence (AI) literacy in teacher education: a scoping review. Comput Educ Open 2024;6:100169. https://doi.org/10.1016/j.caeo.2024.100169.
[121] Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health 2023;5(4):e179–81.
[122] Fu L, Newman B, Jakesch M, Kreps S. Comparing sentence-level suggestions to message-level suggestions in AI-mediated communication. In: Proceedings of the 2023 CHI conference on human factors in computing systems; 2023. p. 1–13.
[123] Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell 2023;5(1):46–57.
[124] Engelbart, D.C. (1962). Augmenting human intellect: a conceptual framework. *Menlo Park*, CA.
[125] Hassani H, Silva ES, Unger S, TajMazinani M, Feely M. Artificial Intelligence (AI) or intelligence augmentation (IA): what is the future? AI 2020;1(2):8.
[126] Echeverria V, Holstein K, Huang J, Sewall J, Rummel N, Aleven V. Exploring human–AI control over dynamic transitions between individual and collaborative learning. In: Proceedings of the *addressing global challenges and quality education: 15th European conference on technology enhanced learning EC-TEL 2020 Heidelberg Germany September 14–18 2020 proceedings*. 15; 2020. p. 230–43.
[127] Holstein K, McLaren BM, Aleven V. Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. J Learn Anal 2019;6 (2). https://doi.org/10.18608/jla.2019.62.3.
[128] Kasepalu R, Prieto LP, Ley T, Chejara P. Teacher artificial intelligence-supported pedagogical actions in collaborative learning coregulation: a wizard-of-oz study. Front Educ (Lausanne) 2022;7:736194.
[129] Law L. Application of generative artificial intelligence (GenAI) in language teaching and learning: a scoping literature review. Comput Educ Open 2024;6: 100174. https://doi.org/10.1016/j.caeo.2024.100174.
[130] Wu CST, Fowler C, Lam WYY, Wong HT, Wong CHM, Loke AY. Parenting approaches and digital technology use of preschool age children in a Chinese community. Ital J Pediatr 2014;40:1–8.