Global Educational Research Journal: 2360-7963
Vol. 13(7) July, Pp. 248-273, 2025

Full Length Paper

Examining the Role of Generative AI in Academic Writing Assessment: A Mixed-Methods Study in Higher Education

[1]Noble Lo and [2]Sumie Chan

[1]Lancaster University, Lancaster
[2]College of Professional and Continuing Education, The Hong Kong Polytechnic University
The Chinese University of Hong Kong, Hong Kong) sumiechan731@gmail.com

Abstract

This study examines the use of generative artificial intelligence (AI) to assist in grading and feedback on undergraduate academic writing. While AI technologies in education have shown considerable potential, their adoption remains controversial due to contrasting evidence about their effectiveness. At a university in Hong Kong, an initiative was launched to use generative AI for providing formative feedback on students' written drafts prior to final submission. This mixed-methods study evaluates the initiative by comparing AI-generated grades with those assigned by human instructors and analysing stakeholders' perceptions of the AI feedback. The research instruments used may partly explain the differences between the quantitative and qualitative research results. Quantitative analysis revealed a high inter-rater reliability between AI-generated scores and human-assigned grades, demonstrating the technical accuracy of the system. However, qualitative data from student interviews and instructor reflections revealed significant concerns about the AI's ability to deliver contextually relevant, personalised, and pedagogically meaningful guidance. Stakeholders expressed scepticism and mistrust, emphasising the lack of depth, specificity, and adaptability in AI feedback as compared to human feedback. By exploring these discrepancies, this study underscores the tension between technical reliability and perceived pedagogical value in AI-assisted assessment. It advocates for stakeholder-informed, hybrid approaches that integrate AI with human oversight to ensure AI tools genuinely support student learning outcomes. To address these challenges effectively, future initiatives should prioritise collaborative design, transparent communication, and stakeholder education.

Keywords: artificial intelligence/English Language Teaching/feedback/grading/higher education, revision

Accepted 16/7/2025 Published 27/7/2025

I. 0. INTRODUCTION

This research aims to critically evaluate the efficacy of generative artificial intelligence (AI) in the context of academic grading and feedback provision. As AI technologies continue to evolve, educators and institutions are increasingly exploring their potential to automate and enhance assessment processes. Prior research, such as Ahmed et al (2023), has highlighted that AI can significantly reduce grading time and improve consistency across assessments. However, questions remain regarding the accuracy, fairness, and pedagogical appropriateness of these systems when applied to complex student responses. This study seeks to contribute to this ongoing discussion by examining the implementation of a specific AI grading and feedback system at a university in Hong Kong, providing localised insights that may inform wider adoption strategies.

The focal point of this study is a particular AI system used to evaluate student assignments at a Hong Kong university. Generative AI models, like those based on large language models, have demonstrated impressive capabilities in understanding and generating human-like text, making them promising tools for educational assessment (Noble et al., 2023). Nonetheless, the extent to which such models can accurately interpret nuanced student responses and deliver meaningful, personalised feedback is still under investigation. By analysing the system's grading accuracy, consistency, and the relevance of its feedback, this research aims to identify both its strengths and limitations. Such insights are crucial for determining whether AI can reliably support educators in the assessment process without compromising fairness or educational integrity.

In addition to evaluating technical performance, this research emphasises understanding the perceptions and attitudes of students and instructors toward the AI grading system. Previous studies, such as those by Wang et al. (2024), have highlighted that user acceptance plays a critical role in the successful integration of AI tools in educational contexts. Exploring how stakeholders perceive the fairness, transparency, and usefulness of AI-generated feedback can reveal potential barriers to acceptance and implementation. To gain deeper insights, the study employs qualitative methods, including interviews and surveys, to capture the experiences, concerns, and suggestions of both students and instructors. Understanding these human factors is essential for designing AI systems that are not only effective but also ethically and socially acceptable.

Ultimately, this research aims to bridge gaps in existing literature by providing a comprehensive analysis of both the technical capabilities and human perceptions of AI-based assessment systems in higher education. While previous research has often focused on either the technological aspects or stakeholder attitudes independently, this study adopts a holistic approach that integrates both perspectives. The findings are expected to inform best practices for deploying AI in academic assessments, ensuring that such systems support fair, transparent, and constructive evaluation processes. By focusing on a real-world application within the context of Hong Kong, this study also contributes valuable insights into the cultural and institutional factors influencing AI adoption in diverse educational environments.

1. 1. Background

This study explores both the efficacy of AI-generated scoring of academic work as well as the perspectives of students on the use of AI towards providing feedback on this work. The extant body of literature suggests that feedback on student work can have significant positive effects with regard to learning outcomes, particularly through supporting students to revise their academic work (Graham et al., 2015; Gnepp et al., 2020). Given this implication, feedback that is of poor quality may affect student outcomes—a concern considering the strain that marking work to provide feedback may place upon instructors who have an already significant workload (Madigan & Kim, 2021; Hahn et al., 2021). There is thus the potential for AI to both reduce the workload upon instructors and to ensure a more standardised quality of feedback for students in the marked work they receive.

The prospect for AI to be applied successfully towards providing students with feedback on their academic work has been posited through research already (Gao et al., 2024; Hooda et al., 2022). Studies suggest that automated feedback supplied by AI systems may constitute feedback of a quality in line with feedback provided by instructors and human markers (Gao et al., 2024; Hooda et al., 2022; Nazaretsky et al., 2024), while others illustrate that computerised, automated feedback can reduce the workload placed upon instructors (Crossley et al., 2022; Machado et al., 2025). The usefulness of AI-generated feedback in providing guidance on written products specifically has been suggested by past research studies (Mertens et al., 2022; Fleckenstein et al., 2023). That being said, there is a general lack of studies that support an empirical basis for asserting the efficacy of these initiatives in terms of student outcomes (Lee et al., 2024).

As university courses move online or take a blended approach to delivery in the post-COVID era, there is increasing opportunity for the implementation of automated, computerised feedback as part of the design of courses and assessments (Ruiz-Palmero et al., 2020). However, concerns about the comparability and accuracy of AI grading may be stalling development and implementation of such programs (Lee et al., 2024; Knoth et al., 2024). This implies the need for further research in this area to close knowledge gaps in the extant body of research literature.

As a consequence of an increasing demand for consistent feedback coupled with the pressures placed upon instructors and supervisors working within contemporary educational environments, there is additionally increasing demand for developing and implementing AI-based feedback systems. This demand is supported by a growing evidential base for their relative efficacy. Within Hong Kong, the government has announced a number of initiatives aimed at supporting the use of digital technologies in education.

1. 2. Rationale

To support such policies and initiatives, it is necessary to both establish how effective such systems are and how they may be best implemented and to monitor their efficacy in terms of outcomes. As is noted above, there is a lack of studies that support an empirical basis for the efficacy of these initiatives and systems. In particular, there is a dearth of research into how accurate and reliable AI grading of written work is, as well as into how far students and instructors respond to specific systems that offer AI grading and feedback instead of or as well as instructor feedback. This gap in the research and the general demand for greater insight into how such systems may be best designed and implemented motivates conducting primary research in this area.

I.3 Aims and Objectives

This study aims at investigating the effectiveness of an automated graded system introduced at a Hong Kong university. It seeks to better understand how effective such systems are for grading and providing feedback on academic work, evaluating the accuracy and reliability of the system in place at a Hong Kong university. This is carried out through comparing the accuracy of the AI grading system with grades assigned by academic staff at the university. In addition to this, the study aims at exploring how students perceive and respond to AI-generated feedback, with a special focus on how they use it to revise written academic work in higher education settings. Through applying these aims, the study is hoped to make a contribution to the design of evidence-based systems for providing AI assessment and feedback, addressing key gaps in the literature identified in the review of extant research below.

1.4 Significance

The research below constitutes an evaluation of a system trialled at a Hong Kong university, undertaken to ascertain how far empirical evidence can support the implementation of AI systems towards providing effective grading and feedback for students. Beyond evaluating the effectiveness of this particular program, the study may likewise contribute to a broader literature on the utility of AI towards providing feedback on student academic work, as is discussed in the below literature review. Through undertaking a mixed-methods approach to research, both the accuracy of AI-based systems as compared with human feedback and assessment may be ascertained, in addition to examining how students view the AI-based system as compared with traditional means of producing and receiving feedback on their work.

2.0. LITERATURE REVIEW

This section reviews the existing research literature on the use of AI in certain aspects of education, focusing on the implementation of automated grading and AI-generated feedback on student work at the university level. It highlights relevant gaps in the literature that this study is designed to close.

2. 1. Automated Grading

Automated grading systems have been in development for much of the twenty-first century, though have previously enjoyed only meagre application beyond trial programmes. A 2021 meta-review of existing research on the automated grading of essays (AGE) noted a number of shortcomings with existing systems, including their comparable accuracy with human grading, as well as widespread negative perceptions among instructors and students alike (Borade & Netak, 2021). The article noted a number of improvements needed to support further uptake, including enhancements to the accuracy of AGE systems, such as in their capacity to assess complex arguments on behalf of students, as well as in their capacity to provide feedback tailored to the individual student.

Past research has echoed many of these sentiments, finding AGE to be unsuitable for grading complex assignments, recommending instead its use on behalf of human markers to reduce workload rather than to replace the human marker altogether (Geigle, et al., 2016). Other studies noted an inflexibility on behalf of AGEs to adapt to different types of task, such as grading diagrams and graphics developed by students (Bian, et al., 2020). Prior to the development of generative AI, AGEs typically had to be coded anew for new assignments, adding also to the cost of their development and implementation in grading and offering feedback on a given module of study. This contributed to limited implementation of AGEs in practice prior to the launch of several generative AI platforms in 2022 (Heaven, 2022).

Since 2023, a number of studies have emerged exploring the application of generative AI programmes utilising large language models (LLMs) towards marking and grading in higher education. A good deal of these studies suggest more accurate grading on behalf of AI systems. For one, a number of studies find that LLM-based AI programmes provide more accurate grades than pre-generative AI AGEs, including those tailored for specific assignments (Schneider, et al., 2024; Xie, et al., 2024; Chu, et al., 2025; Li, et al., 2025). That generative AI programmes such as Chat GPT provide a superior form of grading to tailored AGE systems is now apparently accepted across the literature, transferring research focus onto the potential of LLM-based AI programmes for wider implementation in grading student work.

In addition to this, a number of research studies have compared LLM-based AI grading with that of human markers and found it to be in line with instructor grading. A meta-analysis comparing LLM-based AI models with AGEs found that the latest generation of AI programmes had closed the gap on instructor grading accuracy, rendering them comparably accurate when applied in the assessment of natural-language responses such as essays (Messer, et al., 2023). One study of 120 adult learners on a massive open online course compared Chat GPT grading with both instructor and peer grading, finding it to be on a par with instructor grading and more accurate than peer grading (Impey, et al., 2025). Another study comparing instructor and AI-based feedback on open book examinations found little difference in the grades offered, suggesting its use to reduce workload upon examiners (Dimari, et al., 2024).

However, a number of studies suggest also inaccuracies on behalf of AI grading systems. One study of using Chat GPT to grade Physics papers found that up to 40 per cent of AI responses were inaccurate, recommending that instructors inspect at least 15 per cent of all AI-graded papers to identify the most common areas of inaccuracy (Chen & Wan, 2025). A study of its use in Medical Education also found that GPT provided significantly lower grades than human markers, recommending that its implementation required teacher second-marking for grades below a certain level (Grevisse, 2024). Similarly, a study of the accuracy for GPT to assess coding exercises found significantly lower accuracy as compared with human grading (Lagakis, et al., 2024). The implication of these studies, though they did not directly measure the grading of essays, is that LLMs frequently make factual errors when grading work with binary correct-incorrect responses. It may be, however, that generative AI is better at evaluating the quality of responses wholly expressed in natural language, in line with some of the findings discussed above.

Nevertheless, a number of studies on the use of AI to grade essay responses have found generative AI programmes to be less accurate when compared with instructor marking. One study found that whilst AI grading was more accurate than peer-grading, AI grades when modelled around instructor-provided correct answers was more accurate than those without these model answers, but still was not in line with instructor grading (Golchin, et al., 2024). Another study compared Gemini-pro, GPT-3.5 and GPT-4.0 with instructor feedback, finding the latter to be the most accurate (a common finding across the literature), though still not in line with instructor grading (Lee & Song, 2024). There are therefore mixed findings across the literature with respect to the comparative accuracy of generative AI when it comes to grading essays, inviting more research into this area.

2. 2 AI-Generated Feedback

Research across the twenty-first century has focused on AI's capacity to provide useful feedback for students, either in place of or in addition to instructor feedback. Theoretical approaches to the topic suggest that students may process AI feedback differently from human feedback. For instance, cognitive load theory suggests that an absence of established social cues and shared external referents may impact student understanding of AI-generated feedback (Gonzaga et al., 2025). Assessment literacy literature indicates that students' evaluation of feedback quality can incorporate affective or relational dimensions that transcend technical accuracy (Guo et al., 2025; Zhan & Yan, 2025). In other words, the lack of a human and social element to AI feedback might impact how students receive and respond to it, irrespective of their beliefs about its technical accuracy.

Nevertheless, some theoretical approaches to the topic suggest that AI-generated feedback can provide advantages because of its exclusion of subjectivity in response. One chapter on the future of automated grading systems argues that AI is able both to provide feedback in a more timely—possibly instantaneous—timeframe and to minimise the subjectivity and biases often associated with human feedback (Vetrivel et al., 2025). Thus, AI's ability to escape the personal perspectives of individual instructors should theoretically allow for a more 'universal' view on a student's work and how to improve its quality.

Conversely, other studies have found a tendency for AI feedback to reflect its own biases derived from its training data, suggesting limitations to the usefulness of academic feedback based upon machine learning rather than pre-coded criteria for grading and judgement (Gratani et al., 2024). In particular, the capacity for AI feedback to adequately address the needs of individual learners has been highlighted as a concern. How far AI can address the specific issues of students unknown to the model in the same way as instructors come to know a student has been seized upon as evidence of its limited utility in educational contexts (Lindsay et al., 2025). Such theoretical and ethical concerns underpin much of the contemporary debate about the role of AI in providing feedback on student work.

There is much empirical research that suggests that LLMs have the potential to offer high-quality feedback. One such study had instructors assess both human and AI feedback and found that whilst 'out-of-the-box' LLMs were inadequate, LLMs modified specifically to provide grading and feedback were on a par with human grading and feedback according to independent human raters (Xiao et al., 2025). A similar study employed instructor assessment of peer feedback on argumentative essays, finding that access to an AI chatbot improved the quality of peer feedback (Guo et al., 2025). Other studies on its influence upon or comparison with peer feedback generally find AI to be useful or superior (Lee, 2023; Bauer et al., 2023; Banihashem et al., 2024), though evidence regarding instructor feedback is less established.

However, some studies suggest that AI feedback is not as useful as human feedback when measured empirically (Chan et al., 2024; Lo et al., 2025). Moreover, over-reliance on AI tools led to instances of decreased creativity and critical thinking (한수미 & 김민지, 2025). One literature review of 83 published articles found that whilst perceptions of AI feedback among students and instructors were generally positive, studies measuring the actual impact of feedback upon student performance were few and far between (Shi & Aryadoust, 2024). One such study found that LLM-generated feedback increased revision performance as compared to revising without feedback, linking this to improvements to task motivation as compared with the non-feedback control group (Meyer et al., 2024). Comparisons with receipt of no other feedback generally find that AI feedback improves revision scores (Xu et al., 2023), though again this does not establish how comparably useful AI feedback is with regard to instructor feedback on the same work. Teacher intervention was essential for contextualising AI-generated feedback, guiding students in adapting suggestions, and addressing gaps in communicative and rhetorical skills (Yun, 2025).

Although studies measuring students' performance are scarce, others have indicated clear benefits to affective aspects of students' perceptions associated with meta-learning and assessment performance. A study carried out in China found that AI feedback was associated with significant improvements in motivation, understanding, and preparedness as compared with no intervention (Yeung et al., 2025); however, the study did not compare performance with human feedback. A study on AI feedback and its relationship with student engagement, carried out in Australia, found that instantaneous AI feedback was able to boost student engagement with assessment tasks (Dann et al., 2024). This was not compared with human feedback given that AI provides the possibility for instantaneous, remotely delivered feedback in a way simply not possible on behalf of human instructors. Other studies have likewise found correlations between receiving AI feedback and skills associated with self-regulated learning (Chang et al., 2023; Afzaal et al., 2024). Such studies make the case for augmenting or complementing human feedback with AI feedback but cannot themselves support the replacement of human feedback with AI-generated feedback.

The effect of AI feedback on these skills is likewise not necessarily universal given the mediating effect of student cognition. A study that looked at 6,960 students and another that analysed 8,642 open-ended responses showed that students were unsure about how reliable AI is, seeing its advantages in being easy to access, quick, and providing a lot of information, which they felt were different from the benefits of feedback from teachers (Henderson et al., 2025). A systematic review of student perspectives on AI feedback found significant variation across both national cultures and levels of education (Atherton et al., 2024), suggesting the utility of research carried out within specific cultures and institutions.

2.3. Theoretical Framework

This section sets out the theoretical framework behind this study, incorporating models and perspectives influence by Theory of Change, participatory evaluation, and Rationale, Uses, Focus, Data, Audience, Timing, Agency (RUFDATA). These frameworks provide a lens through which the processes of learning in response to AI-generated material may be explored.

2.3.1 Theory of Change (ToC)

In order to evaluate the efficacy of an AI system, it may be helpful to consider what theoretical approach has the capacity to ascertain how far the initiative investigated meets its aims. Theory of Change (ToC) is capable of providing a framework for outlining how and why an educational outcome may be expected to occur within a given educational context (Reinholz & Andrews, 2020). Articulating a theory of change requires both identifying the requirements of the initiative with respect to what the intervention is intended to address – for example, reducing teacher workload whilst at the same time ensuring consistent feedback – as well as exploring the mechanisms that underpin how the design of the programme's core components function to elicit change. Through following such a process, outcomes that are measurable can thus be arrived at so that the efficacy of the programme – in this case, the AI based marking and feedback system – may be adequately tested. Taking such an approach to this research study, the initiative can be evaluated in terms of its stated goals and how their assumed outcomes are met and measured (Davies, 2018). Undertaking such an evaluation thus requires undertaking research that can measure the consistency of AI generated grading in comparison with the lecturer marking.

A further element that is part of ToC related to the AI system is with respect to student responses to the initiative. Whilst marking papers accurately may reduce teacher workload without impacting the accuracy of grades given, a considerable part of the grading process often involves giving feedback on draft and final submissions. The relationship between useful feedback and student improvement is well attested according to the literature on this topic (Graham et al., 2015). This evidence implies the importance of taking an approach to evaluating the efficacy of an AI system that also evaluates student experience of the feedback given, so that its relationship with student outcomes may be properly explored.

ToC thus not only serves as a means to better understand the aims of the AI grading system in terms of promoting change in education settings, but also can assist in determining which outcomes ought to be tracked in aid of measuring change. The extent to which the AI programme can ensure consistent marking, provide high-quality feedback, and reduce student workload thus serve as three measures of change consistent with a successful programme or intervention in accordance with the review above.

2.4 Research Gap

A number of gaps in the research literature have been outlined in the review above. First, while LLMs have apparently improved grading accuracy as compared with previous generation AGEs, there are mixed results with regard to their accuracy as compared with human markers, especially for essay-based tasks. Whilst LLMs perform better in assessing natural language responses as compared with the binary responses associated with some STEM subjects, the presence of factual inaccuracies in the latter raises questions about the accuracy of grading when it comes to essay-based questions, many of which may contain factual assertions. Many studies only recommend AI grading within a context of human oversight models in which humans remain 'in-the-loop' when it comes to automated grading processes. There is therefore research required to establish the accuracy of AI when it comes to grading essay-based work as compared with human markers in specific educational contexts.

The research literature on AI-generated feedback also indicates significant gaps in present understanding. Few studies have directly compared student learning outcomes because of AI-generated feedback versus those influenced by human instructor feedback. There remain open questions about the capacity of AI feedback to replace or 'stand in' for instructor feedback. Central to the debate are questions of biases in AI feedback derived from training data and how individualised AI feedback can prove in comparison to instructor feedback. Though many studies indicate positive relationships between AI feedback and student experiences, extending in some cases to its empirical impact on revisions made to student work, these findings are overwhelmingly derived from comparisons with peer feedback or with no feedback at all, rather than with instructor feedback. Likewise, the literature suggests that cultural and individual perspectives may serve to mediate the perceived efficacy of AI feedback, indicating that localised and context-specific research may be required to establish its utility within a given educational context.

2 5. Research Questions

The aim of this study is to investigate both the precision of the automated grading system introduced at a Hong Kong university through comparing marks assigned to work by generative AI with work graded by instructors at a Hong Kong university. In addition to this, the study investigates the student perspective on and experience of the feedback received on AI graded work. These aims may be refined into the following research questions:

- The comparison of accuracy of grading of students' work by human instructors and AI based system and addresses research gap in the existing literature.
- How do students experience and respond to AI generated feedback?
- Answering these questions will require undertaking mixed-methods research consisting of both qualitative and quantitative approaches. The design of this is set out in some depth in the chapter on methodology below.
- In addition to this, and in recognition that the implementation of new systems takes place within complex educational environments, this study is aimed at exploring the foundational theoretical constructs pertaining to automated evaluation of work within in academic settings, which is reflected in the study's theoretical framework and its discussion. This aim may be formulated through the following research question:
- How can AI based feedback systems be effectively implemented from the perspective of Theories of Change (ToC) and Participatory Evaluation?
- The theoretical approach taken in pursuit of this research is elaborated upon in the relevant section below.

2.6. Participatory Evaluation

The utility of including student perspectives in this study is implied also by a theoretical framework informed by participatory evaluation. Participatory evaluation maintains that stakeholders within a given environment ought to be actively engaged in any evaluation process that involves them (Makgamatha, 2009). Whilst exploring the impact of AI-based feedback on students's written work forms the focus of forthcoming research (Lo et al., 2024), investigating student evaluations of an AI-based system in qualitative terms (e.g., experiences, attitudes, etc.) reflects an evaluative process rooted in participatory evaluation, recognising the validity of student experiences as stakeholders within the initiative.

Taking a participatory-evaluative approach is anticipated to inculcate a sense of ownership over and promote engagement with the initiative, which in turn is hoped to contribute towards the future design and implementation of effective initiatives for change within their educational environment. In terms of research design, this has informed the inclusion and design of qualitative interviews insofar as the perspective frames both students and instructors as stakeholders whose insights are required to ascertain the impact of the trialled system. Gathering these experiences and perspectives allows for the research to take into account the human element of participation in educational systems, helping to identify benefits or disadvantages undetectable in statistical analysis that are nonetheless impactful with respect to relevant outcomes.

2.8. RUF Data

To meet the study’s objectives, this design of the research was informed by RUFDATA, a practical framework for planning evaluations (Saunders, 2000). This approach is designed to guide the evaluation planning process and can be informed by and integrated with ToC and participatory evaluation throughout its stages of planning. Accordingly, the planning of the evaluation process responds to prompts across the following areas:

Table 1: RUF Data Research Design

Rationale	Why conduct the evaluation?
Uses	What is the anticipated utility or applications of the evaluation’s findings?
Focus	What is the main focus of the evaluation?
Data	What data will be collected and what purpose will this serve?
Audience	Who is the intended audience for the evaluation?
Timing	When will the evaluation take place?
Agency	Who shall conduct the evaluation?

This framework thus provides a structured approach to aligning the study’s design with its practical aims, justifying in this case the inclusion of both qualitative and quantitative arms to the research carried out. In applying this model for planning evaluation—informed both by ToC and by participatory evaluation—an effective and inclusive design for this evaluation of the initiative may be arrived at through taking into consideration the rationale, data needs, and audience for the research. How this is achieved is described and justified in the section that follows.

3.0. METHODOLOGY

This section presents the research design behind this study, justifying its methodological decisions in terms of the demands of the study’s research questions and theoretical framework.

3. 1. Research Design

This study is situated within a social constructionist framework. Social constructionism holds that social practices are influenced by the norms of cultural institutions (Jung, 2019), whilst at the same time viewing social norms as resultant from social practices (Witkin, 2012). This implies the necessity of an interpretivist approach to research, according to which researchers must attempt to understand the beliefs and reasoning that shape the behaviour of social actors (Pulla & Carter, 2018). For these reasons, research methods that are narrowly positivist—admitting only quantitative analysis of material events (Su, 2018)—will fall short of the requirement to understand how relationships between social phenomena are mediated by human beings as thinking agents (Gergen, 2015). When constructing a research design within such a framework, it is therefore important to include qualitative methods that can provide an explanatory account for the mechanisms behind observed statistical correlations or trends (Ivankova et al., 2006).

3. 2. Context and Approach

The initiative introduced at the university utilises AI generated marking of and feedback upon student work as part of various modules in which written work was submitted by university students. The AI system employed provides grades and feedback for draft versions of papers provided by students, before their final papers are graded by instructors. The marks assigned by the AI system will be compared with those proffered by human markers on the same piece of draft work so as to assess the accuracy of grades generated by AI. This may be achieved through employing statistical analysis of the marks assigned by both parties, establishing whether AI given grades are generally consistent with those given by human markers. To this end, several instructor marks for the same work will need to be compared with a mark given by AI. In the design described below, draft work is triple marked by different human instructors so that a baseline for instructor grades may be produced against which AI marks can be compared.

As mentioned above, the research also explores student perspectives on AI generated feedback and as such employs qualitative analysis of data collected from students. This reflects a mixed-methods design that triangulates the findings of quantitative analysis with that of qualitative analysis, following an explanatory design that seeks to understand observed statistical trends in light of stakeholder perspectives and experiences (Ivankova et al., 2006). This is to be accomplished through holding interviews with students who have participated in courses where the initiative has been applied and who have thus received AI feedback. These interviews should provide a suitable means for investigating student perspectives in that they can be used to explore what students think about AI feedback but also can be used to establish *why* they have formed these opinions based on their experience (Bolderston, 2012). Interviews can also allow for personal perspectives to be communicated in a way that is detailed as compared with alternative means such as questionnaire data (Bazeley, 2013). Taking this approach can thus help fulfil the aims of inviting the participation of key stakeholders in the process of evaluation.

3. 3. Data Collection

So as to arrive at a statistically meaningful analysis of the quantitative data, the research design involves the analysis of marks awarded to 150 different papers submitted as part of the initiative. The papers in question involved essay questions requiring responses of no more than 1000 words, all of which were written in English in accordance with the standards of the ESL programme. Students submitted work digitally via the usual avenues and then were distributed via email to human markers, whilst the researcher was responsible for uploading the papers to AI for grading and feedback. These initial papers were first marked by AI and then marked again by three different markers, none of whom are aware of the AI marks given, nor that of each other. All human markers were supplied with a rubric for the papers and marking guide, whilst the AI receives a comprehensive prompt outlining learning objectives. Through triple marking each paper, a collection of marks assigned by human markers was established so that AI grades may be compared against them. This data was recorded in a database ahead of quantitative analysis.

In terms of the sampling of student participants, ten participants were deemed as sufficient for undertaking a small-scale study with thematic analysis according to published guidance on research methods (Hammersley, 2015). Students were provided with the option of participation and then interested parties randomly selected using purposive sampling (Palinkas et al., 2016), with the process of selection filtering out those who had not read or engaged with the AI generated feedback. Those selected were interviewed in a one-to-one, face-to-face setting and were each posed ten interview questions pertaining to their experience of receiving AI generated feedback as part of their education (appendix 1). In line with Allen (2017), the interview questions were designed to be open-ended in order to encourage a more detailed set of responses. Additionally, they were posed according to a semi-structured approach so as to ensure a focus on the research topic whilst also allowing the researcher to prompt for detail on areas of particular relevance (Zhang & Wildemuth, 2017). Interviews lasted no longer than an hour and were carried out by the same researcher, who recorded responses using digital recording software. Recordings were stored in password-protected files in line with data protection standards (Resnik, 2020) .

In order to meet the demands of participatory evaluation, instructors' views were also solicited in recognition of their status as stakeholders in the initiative. Instructors were asked to provide commentary on the feedback on student essays generated by AI, using marginal comments on word documents to make remark upon the text. They were asked to evaluate the effectiveness and accuracy of AI feedback on student work in order to add perspective on the strengths and weaknesses of AI generated feedback. The instructors recruited into the study to mark the student essays were thus asked to complete these evaluations, though not on any work they themselves had personally marked. Instructors were provided with a separate set of prompts to that of students, being asked to leave remarks on AI feedback presented to students rather than being interviewed themselves.

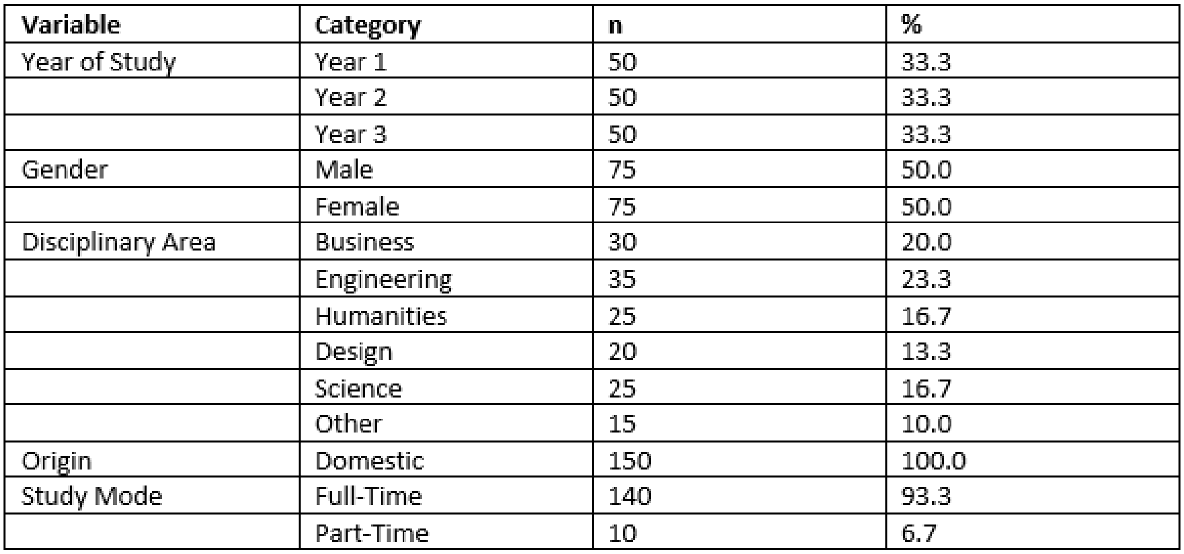
The study also incorporates a sample of 150 student papers produced at the drafting stage of the initiative. The study involved a random sample of 150 papers derived from the many submitted over the course of the initiative. All papers were under 1,000 words in length – to accommodate the limitations of data-handling on behalf of generative AI – and were derived from a number of subjects, though all submitted by students at the undergraduate level. The procedures involved were, from the students' perspectives, identical to that of their regular procedures for submission of formal written work and all were presented as counting towards assessment. Students were informed that their work would be graded in accordance with AI as part of an experiment, but were notified that only instructor-given grades would count towards their degree, in line with the conditions for the school-wide initiative.

Of the roughly 100 instructors involved in the initiative, nine participants were drafted into the interview study – again, through purposive sampling methods (Palinkas et al., 2016). Three of these instructors were selected to mark the essays and to provide their feedback on the initiative following the placement of an advertisement through the university's internal communications network. All human participants were provided with details regarding how their data may be used and were informed of their right to withdraw at any time.

3. 4. Participants

The 300 papers marked in total (150 initial drafts and 150 revised drafts) were drawn from 150 undergraduate students randomly sampled from a cohort. Stratified random sample was employed to ensure that various subgroups were represented accurately. A breakdown of these statistics is conveyed in Table 2. All students were participating in a compulsory English as a Second Language course (ESL), though were drawn from a number of different disciplinary majors. All participants were domestic students to avoid accounts affected by differing native languages given the ESL context of the assessment task.

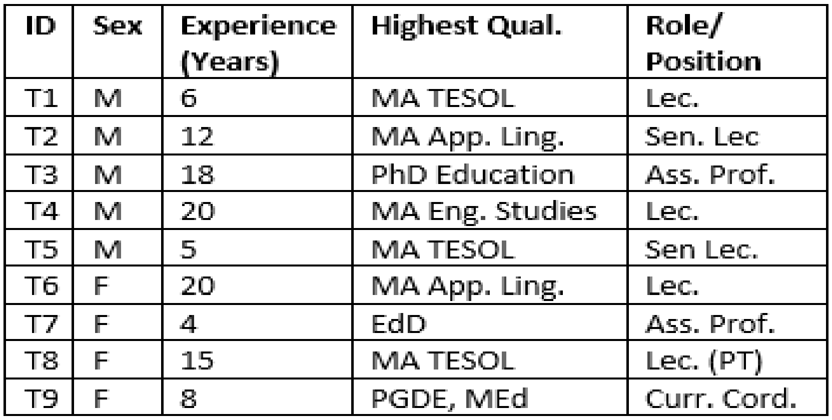
Table 2: Student participant's variables in quantitative grading study



Variable	Category	n	%
Year of Study	Year 1	50	33.3
	Year 2	50	33.3
	Year 3	50	33.3
Gender	Male	75	50.0
	Female	75	50.0
Disciplinary Area	Business	30	20.0
	Engineering	35	23.3
	Humanities	25	16.7
	Design	20	13.3
	Science	25	16.7
	Other	15	10.0
Origin	Domestic	150	100.0
Study Mode	Full-Time	140	93.3
	Part-Time	10	6.7

The students participating in the interview study were also sampled purposively to be as representative as possible, though naturally this was not wholly possible given the smaller cohorts involved in the interview study. As Tables 2 and 3 demonstrate, all students participating were full-time students, whilst all but one instructor were at full-time staff. Among the instructors, there was some variation in their highest level of qualification, whilst their level of seniority also varied somewhat.

Table 3: Instructor participant variables in instructor interview study



ID	Sex	Experience (Years)	Highest Qual.	Role/ Position
T1	M	6	MA TESOL	Lec.
T2	M	12	MA App. Ling.	Sen. Lec
T3	M	18	PhD Education	Ass. Prof.
T4	M	20	MA Eng. Studies	Lec.
T5	M	5	MA TESOL	Sen Lec.
T6	F	20	MA App. Ling.	Lec.
T7	F	4	EdD	Ass. Prof.
T8	F	15	MA TESOL	Lec. (PT)
T9	F	8	PGDE, MEd	Curr. Cord.

3. 5. Data Analysis

Analysis of the interview data utilised qualitative rather than quantitative methods, employing thematic analysis to this end. Thematic analysis generates representative 'themes' from a dataset to allow for particularly prevalent and emphatic themes to be brought to the fore (Braun & Clarke, 2006), providing both findings representative of participant perspectives and also generating a useful structure for exploring the specific responses of individual participants (Evans, 2018). Thematic analysis also aligns well with the study's theoretical framework set out above due to its theoretical freedom (Nowell et al., 2017). Through searching for themes that describe the phenomenon under investigation, thematic analysis holds the capacity to demonstrate in a clear way what it is that students believe about receiving marks and feedback generated by AI.

This outcome was achieved in the study through carrying out analysis in Leximancer. This application was selected because of its capacity to use preset algorithms to extract semantic and relational data from texts and to aggregate that data into themes (Smith & Humphreys, 2006). This application functions by representing connections between conceptual terms as lists ranked by their prevalence. The algorithm can show which concepts are most prominent, related, or co-occurring. Through this approach to analysis, student and instructor perspectives may be analysed in a systematic manner, and they can then be further illustrated through excerpts from the interview texts themselves. This analysis produced several major themes that are used to structure discussion of the interviews themselves below.

Following completion of thematic analysis in Leximancer, further analysis of the interview data was carried out to establish whether participation evaluations were positive or negative. This approach addresses a limitation of thematic analysis alone, as the standard Leximancer algorithm only reports on conceptual correlation and frequency without providing information about participant evaluations of these concepts. To establish whether respondent sentiments were positive or negative, student responses and instructor remarks related to the most prevalent themes were manually coded as either positive, negative, or neutral to establish the general sentiment surrounding these concepts. A subset of 20 percent of all comments relating to the most prevalent concepts were coded in this way, permitting the generation of descriptive statistics reported in Tables 5 and 11.

Analysing the marked papers was undertaken using IBM's Statistical Package for the Social Sciences (SPSS) 29.0. Using this program allowed for a variety of tests to be carried out on the same dataset (Salcedo & McCormick, 2020). Tests were carried out to examine how far the AI grading fell in line with the grades given by human markers, thus establishing its accuracy against a human average. These tests sought to establish whether the marks provided by AI generally fall within the trends implied by instructor grades to a statistically meaningful extent. A number of tests were carried out, such as calculating Spearman's correlation coefficient to establish whether instructors and AI ranked essays similarly, as well as the intraclass correlation coefficient to establish inter-rater reliability. An alpha of 0.05 was applied to Spearman's coefficient, and 95% confidence intervals were calculated for the intraclass coefficient.

First, Spearman's rho was carried out to examine the rank-order correlation between the AI-assigned marks and those of human markers. This statistical measure is able to assess the strength and direction of an association between two ranked variables (Daniel, 1990), meaning that it is the rank that is compared as opposed to the raw score for each essay. In the first test, an average between the human-given grades was taken to provide a single comparator score for each individual essay. However, in order to address any problems caused by averaging—such as its masking of differences in individual variation—Spearman's rho was again calculated between the AI grades and those of each of the three human graders.

To triangulate these findings, the intraclass correlation coefficient (ICC) was calculated to establish inter-rater reliability (IRR). An ICC is useful when measuring agreement across more than one rater insofar as it can be used to detect and exclude outliers (Liljequist et al., 2019). Two ICCs were arrived at by using a two-way random-effects model with single measures and absolute agreement so as to facilitate generalisability as well as remain sensitive to discrepancies in scoring. The ICC for the grades assigned by all three human markers was first established, and then a further ICC was carried out, including the AI-assigned grades in the test. Through these means, alignment between the scores given by each marker generally could be established to test whether the AI marker was generally in alignment with the human markers..

3. 6. Ethical Considerations

Several ethical measures were implemented to ensure the integrity and confidentiality of the research process. Informed consent was obtained prior to participation in the study, clarifying how data would be used and outlining the rights of participants to withdraw from the research (Oliver, 2010). Power dynamics were carefully considered and managed to avoid introducing social desirability bias into participant responses (Qin, 2016). Furthermore, the data generated by students was kept confidential and not made subject to dissemination beyond the researchers and markers involved in the study, in keeping with standards for data protection in research (Mourby et al., 2019). Students and instructors participating in the study likewise had their identities anonymised as far as possible to encourage them to speak freely regarding their experiences of the initiative (Saunders et al., 2015). The inclusion of instructors and students as stakeholders fulfils the requirements of participatory evaluation, given their role in assessing the accuracy and efficacy of the initiative through their contributions to the qualitative data in this study.

4. FINDINGS

This section presents the findings to the quantitative and qualitative methods of analysis, beginning with statistical analysis of inter-rater reliability and then presenting the findings from the Leximancer analysis of conceptual themes across the interviews. It concludes with discussion of the findings approached from the perspective of Theories of Change.

4. 1. Quantitative Analysis

4. 1 1 Interviews with Students

The interviews with 10 students who had participated in the initiative were transcribed and then uploaded for analysis via Leximancer. Five main themes were developed through this process, as presented in Figure 1. These themes are discussed in more depth with relation to the analysis findings.

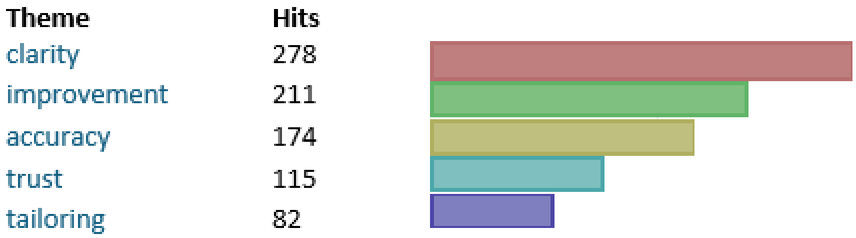


Figure 1: Major themes derived from student interviews with illustration of conceptual frequency

Table 4: Sentiment-based content analysis for main concepts/themes derived from student interviews

Theme	Total Mentions	Mentions Coded (20%)	Positive (%)	Negative (%)	Neutral (%)
Clarity	278	56	16 (29%)	31 (55%)	9 (16%)
Improvement	211	42	14 (33%)	21 (50%)	7 (17%)
Accuracy	174	35	7 (20%)	24 (69%)	4 (11%)
Trust	115	23	4 (17%)	17 (74%)	2 (9%)
Tailoring	82	16	2 (13%)	12 (75%)	2 (12%)

Through utilising a medium detail level of analysis, Leximancer was used to identify which concepts were most associated with the main themes derived from the analysis of the interviews. These were determined on the basis of frequency and proximity of terms to various themes alongside the use of the concept pathway feature that links certain terms to themes across the interviews.

As Table 4 demonstrates, the main themes derived from this approach to analysis were subject to largely negative evaluations of the concepts in question. We looked at a sample of the comments linked to each of the five main ideas and found that 50 percent or more of the responses were negative. Roughly three-quarters of remarks regarding the themes of trust and tailoring were negative, though roughly a third of comments regarding clarity and improvement were positive. These sentiments are reflected in the discussions of each theme below.

In order to examine how far the marks for student essay papers generated by AI fit within the broader trends of marks awarded by human markers, inter-rater reliability (IRR) was calculated through several means. First, Spearman's rho correlation coefficient was calculated by comparing the rankings of AI marks with those of human markers. Spearman's coefficient is a useful means for comparing between two sets of numerical scores, though it is limited to comparisons between two sets (Spearman, 1904). For this reason, an average of the human scores for essays was taken and compared with the scores generated by AI for the same papers. As Table 7 demonstrates, the test generated a coefficient with a strong positive correlation (rs = 0.895) that fell well within the alpha for statistical significance (p = 7.16288E-54). This would suggest that AI marks are very close to those of human markers.

However, there are several problems with this interpretation. For one, using an average score for the human markers may hide outlying results, assuming equal reliability among the human markers. For this reason, we undertook further tests comparing the AI grade with that of each human marker in turn. Results demonstrate a similarly high level of agreement between ranks and strong confidence in the statistical significance of the results. When comparing the AI marker with the first human marker (hereafter, H1), there was a coefficient of 0.811, indicating a high level of agreement between the marks awarded. Similar results were seen when comparing the AI with H2 and H3, showing strong agreement (rs = 0.873 & rs = 0.886), and all three tests had very low p-values, indicating they are statistically significant (p = 2.79712 × 10-36, p = 6.90347 × 10-48, p = 3.17704 × 10-51). This suggests that the AI agreed with some human markers more so than they did with each other, with H1 generating scores a further distance from H2 and H3 as compared with between those markers and the AI.

A different means for calculating IRR was pursued in order to triangulate the findings. The intraclass correlation coefficient (ICC) was calculated to measure the agreement between the various raters involved in the study. The ICC is particularly useful when it comes to comparing more than one rater, to infer generalisability beyond the population admitted to the study, and to detect if any one rater is at odds with the rest (Koch, 1982). For this reason, two ICCs were calculated: one including the AI marker and one without. This allowed for IRR to be calculated for the human markers together and with the AI, with a large difference between these two tests (e.g., a substantial drop in correlation coefficient) suggesting that the AI marks do not fit with the other raters as well as they do with each other.

A two-way random effects model was used in order for the results to be generalisable beyond the population, whereas single measures were selected insofar as the scores taken for each marker were not averaged. Absolute agreement was defined in order to factor in errors and outlying scores made by raters, allowing for inconsistent results to be counted in the analysis. When calculating the ICC for the three human markers, there was an ICC of 0.892, falling within the 95% confidence intervals of 0.87 and 0.915. This indicates a very good level of IRR between the three human markers. A significant decline in ICC for a calculation including the AI marker might indicate a low IRR between it and human markers. However, the ICC for the papers marked by the AI and human markers was 0.88—a decline of only 0.012—which also fell within the 95% confidence intervals of 0.87 and 0.915. The AI marker thus can be established to be a reliable marker on the basis of two different IRR tests.

4. 2. Qualitative Analysis

4. 2.1 Clarity

The theme of clarity was associated with several terms, the most prevalent of which were ‘understanding’ and ‘comprehension’ (Table 5). As the interviews reveal, these were often employed in a critical fashion with respect to the utility of AI: “Sometimes I feel that the program doesn’t understand what I’m actually writing.” (Student 1 [S1])

“It [the generative AI feedback] will at times offer advice for the wrong stuff, like teacher comments on my work, or act like it is marking my work for a different question or subject.” (S4)

The participants were not wholly negative about AI in this respect, being somewhat mixed in their evaluation of its capacity to provide ‘instructions.’ For instance, some celebrated the ‘simplicity’ of the guidance offered, whereas others felt that the feedback was not sufficient in ‘detail’:

“So one thing I like is that it puts forward what you need to accomplish in a very simple way. It doesn’t get lost in jargon and can illustrate its criticisms with examples.” (S10)

“I don’t find it useful. It frequently lacks clarity and appears to be taking a risk when providing suggestions for enhancement. It will say ‘correct your references’ but not actually give examples as to what is wrong.” (S2)

This reflects student criticism of the AI feedback as being insufficiently detailed in its ‘explanation’ as compared with their instructors or tutors.

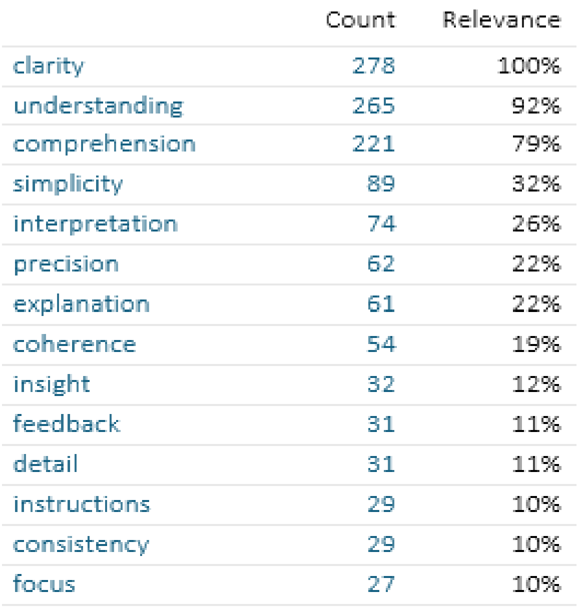
It may be that the focus on AI feedback in the interview questions and instructions provided to the marking instructors may have skewed the discussion towards assessments of AI feedback. Past research has suggested that both instructors and students are fairly negative about the quality of AI feedback (Tishenina, 2024), and some studies have questioned whether it is comparable in quality to that of instructor feedback (Celik et al., 2022). However, it is notable that some students still questioned the accuracy of AI grading of their essays:

“The marks I received I don’t feel are reflective of what an instructor would produce. My lecturers would give me much higher marks than the AI did. (S8)

“It was all over the place. I was getting told it was a first, then, after revisions, a 2:2. Ridiculous.” (S5)

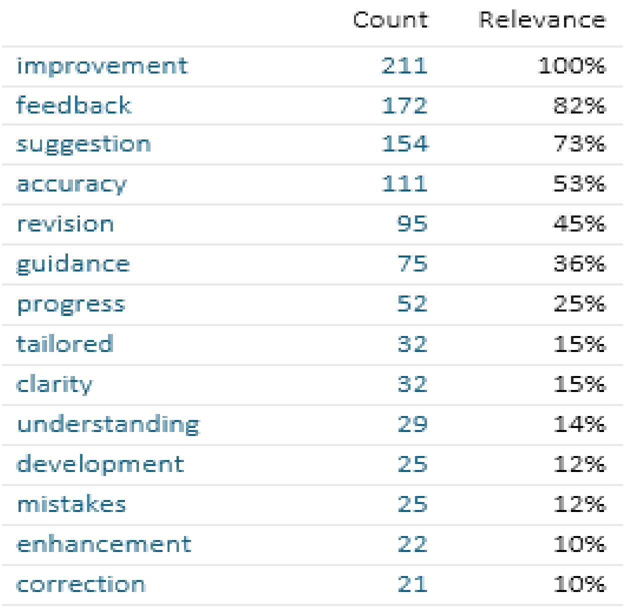
Although the AI grading may have produced some outlying results, this was clearly not frequent enough in the sample analysed to set it apart from the human markers, who evinced a lesser but noticeable degree of personal variation at odds with the group average. Student perceptions may therefore be interpreted in terms of views and experiences of AI that are subject to other influences.

Table 5: Ranked concepts associated with the theme of ‘clarity’ derived from student interviews



	Count	Relevance
clarity	278	100%
understanding	265	92%
comprehension	221	79%
simplicity	89	32%
interpretation	74	26%
precision	62	22%
explanation	61	22%
coherence	54	19%
insight	32	12%
feedback	31	11%
detail	31	11%
instructions	29	10%
consistency	29	10%
focus	27	10%

Table 6: Ranked concepts associated with the theme of ‘improvement’ derived from student interviews



	Count	Relevance
improvement	211	100%
feedback	172	82%
suggestion	154	73%
accuracy	111	53%
revision	95	45%
guidance	75	36%
progress	52	25%
tailored	32	15%
clarity	32	15%
understanding	29	14%
development	25	12%
mistakes	25	12%
enhancement	22	10%
correction	21	10%

Student and instructor bias against the use of AI in educational settings has been well-documented across the literature (Nazaretsky et al., 2024). It may be that students have a negative bias against AI that coloured their experience of it when receiving grades for their papers. Alternatively, it may be that the sample of 10 students interviewed was not sufficiently representative of the student body. Doing a statistical analysis of how students and teachers feel about using AI for grading could help clarify their opinions on this part of the initiative.

4.2.2 Improvement

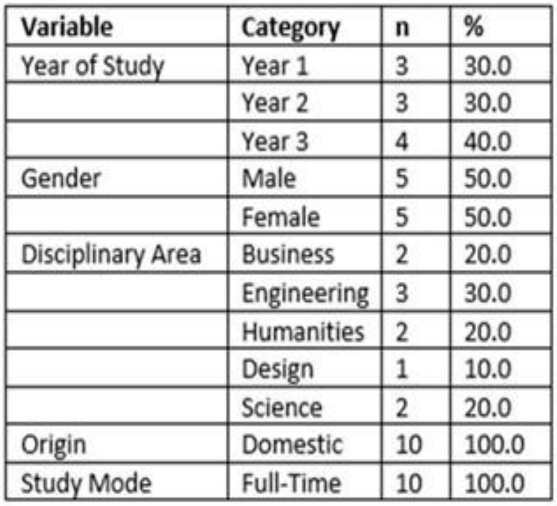
Leximancer identified a number of terms used in conjunction with discussions about ‘improvement’ (Table 6), including feedback (82%), suggestion (73%), and accuracy (53%). These discussions were largely centred around AI’s capacity to make suggestions for revisions capable of improving the student’s performance and grade. Students were mixed with respect to the capacity of AI to complete accurate ‘grading’ of their draft work:

“I was actually surprised that the grades on my draft work were generally correct. It said they would be 2:1s, and they were.” (S10)

“It was all over the place. I was getting told it was a first, then, after revisions, a 2:2. Ridiculous.” (S5)

There was some agreement that AI could generally provide ‘guidance’ capable of informing useful ‘revision,’ though only one participant stated outright that they felt such assistance was better than that which an instructor could provide. One complaint in particular was that AI failed to make note of progress between one draft and the next, often giving advice that contradicted the last guidance received, with one participant comparing it to being marked by different instructors or one instructor with a poor memory. There was some scepticism, then, on how far AI could be used to help students improve their work over time on par with the opportunities for improvement offered by instructor feedback.

Table 7: Student participant variables in student interview study



Variable	Category	n	%
Year of Study	Year 1	3	30.0
	Year 2	3	30.0
	Year 3	4	40.0
Gender	Male	5	50.0
	Female	5	50.0
Disciplinary Area	Business	2	20.0
	Engineering	3	30.0
	Humanities	2	20.0
	Design	1	10.0
	Science	2	20.0
Origin	Domestic	10	100.0
Study Mode	Full-Time	10	100.0

4. 2.3 Accuracy

Terms associated with the theme of accuracy include precision, mistakes/errors, consistency and relevance (Table 8). A number of the respondents felt that the feedback from AI lacked 'accuracy', was 'inconsistent' and lacked 'consistency':

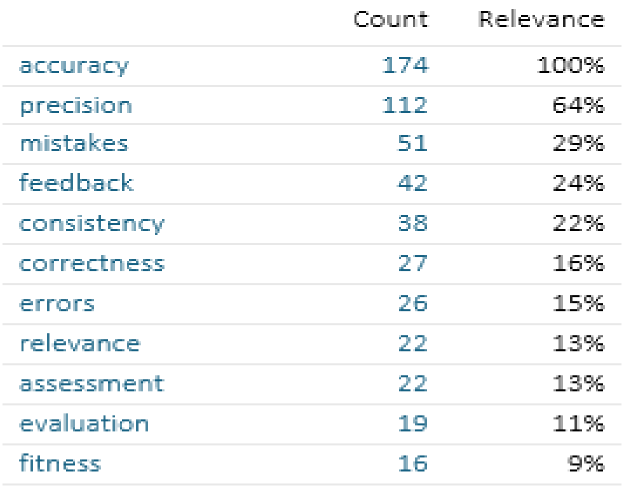
"Some of the guidance given to me by the AI was not accurate […] It didn't correctly grasp the purpose of the essay and misunderstood content." (S7)

"I noted some of the feedback made factual errors when offering guidance…" (S3)

"Some aspects to the feedback such as picking up on grammatical errors were great. But there was no consistency. There were parts it picked up on and parts it didn't." (S9)

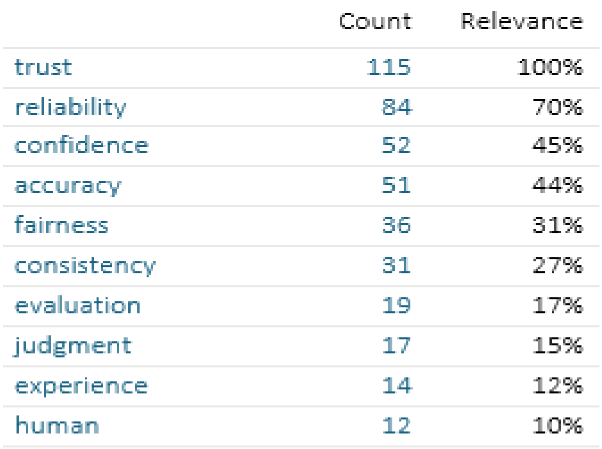
The students interviewed also challenged the 'relevance' of much of the AI's feedback. One student found that the programme focused on aspects to the essay that were tangential or unlikely to improve the grade, whereas another noted that guidance about improving critical evaluation ignored the descriptive nature of the task. Some participants felt that the feedback was not adequately relevant to the 'assessment' rubrics and marking criteria, instead offering general rather than specific feedback.

Table 8: Ranked concepts associated with the theme of 'accuracy' derived from student interviews



	Count	Relevance
accuracy	174	100%
precision	112	64%
mistakes	51	29%
feedback	42	24%
consistency	38	22%
correctness	27	16%
errors	26	15%
relevance	22	13%
assessment	22	13%
evaluation	19	11%
fitness	16	9%

Table 9: Ranked concepts associated with the theme of 'trust' derived from student interviews



	Count	Relevance
trust	115	100%
reliability	84	70%
confidence	52	45%
accuracy	51	44%
fairness	36	31%
consistency	31	27%
evaluation	19	17%
judgment	17	15%
experience	14	12%
human	12	10%

4. 2.4 Trust

The participants mentioned numerous terms in relation to trust, such as reliability, confidence, accuracy, and fairness (Table 9). The word 'trust' was again used largely negatively with respect to feedback:

"I don't really trust AI to offer me good feedback." (S1)

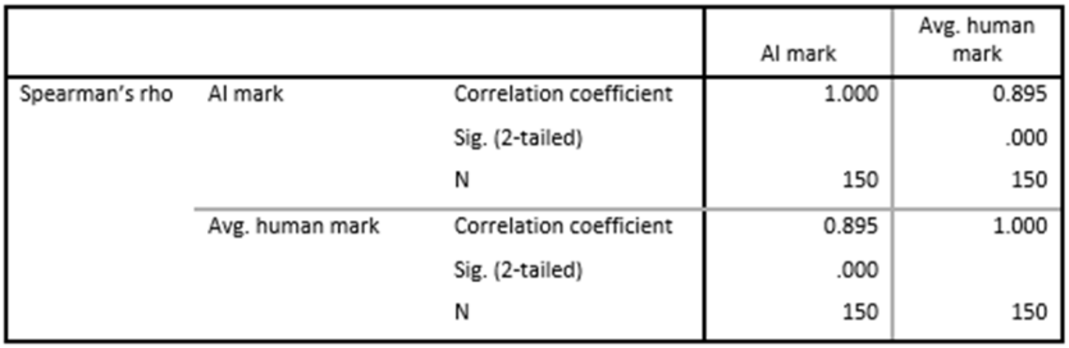
"I wouldn't trust it as much I'd trust something my supervisor told me." (S4)

Participants questioned how 'reliable' AI was at providing consistently useful feedback and accurate grades, noting a great deal of difference between one response and the next. An experience of the feedback as being 'unfair' was inferred from the interviews, related to the inconsistency of the feedback and grading:

"I think one of the great ideas about AI is that it should be objective but I can't see how it can look at a near identical essay twice and give two totally different evaluations." (S9)

The term 'human' was raised largely in conjunction with the word 'trust' and in a comparatively positive sense. Participants reported trusting the feedback and marks of human instructors more so than AI iterations, reflecting a broad distrust of AI technologies.

Table 10: Spearman's rho correlation coefficient for AI marks and average marks of human markers



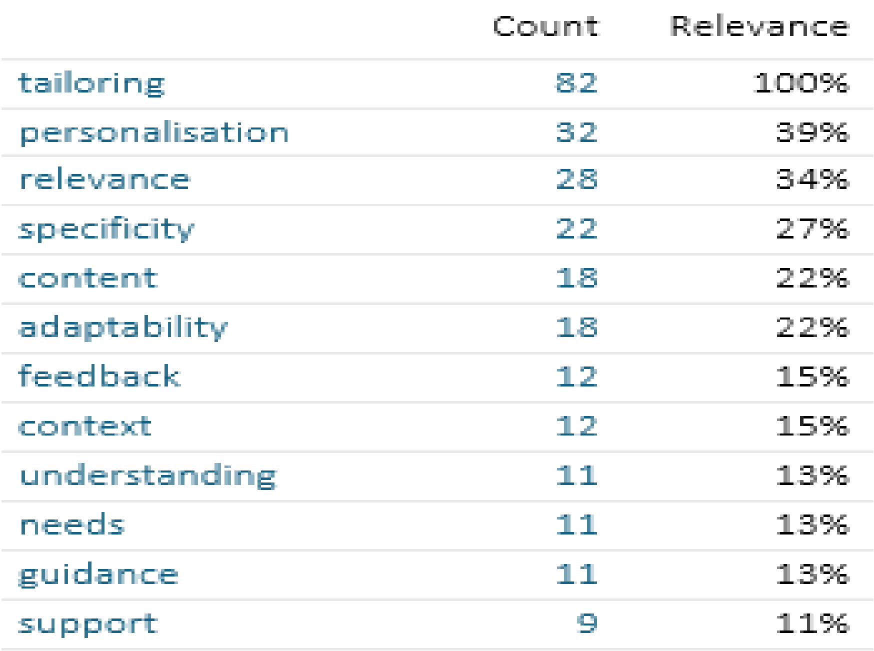
			AI mark	Avg. human mark
Spearman's rho	AI mark	Correlation coefficient	1.000	0.895
		Sig. (2-tailed)		.000
		N	150	150
	Avg. human mark	Correlation coefficient	0.895	1.000
		Sig. (2-tailed)	.000	
		N	150	150

4.2.6 Tailoring

Table 11 illustrates some of the correspondence of various terms with the theme of 'tailoring'. This theme was with response to the extent to which AI feedback was tailored to the student and their work. The students were largely sceptical of the capacity for AI to tailor its responses to them as students based on their experience: "The programme doesn't remember or know anything about me. It can't discover that English isn't my first language. It doesn't know that my second draft built on my first draft. It isn't tailored to me." (S8)

There was a lack of 'personalisation' in the AI feedback that mirrors the lack of 'specificity' in responses. Whilst the feedback was tailored to the essay itself, the students felt that it often felt like general advice illustrated through excerpts from the essay. How 'adaptable' AI could be to different types of assignments was also raised, with one student noting that AI was not able to correctly assess an image attached to an essay they submitted.

Table 11: Ranked concepts associated with the theme of 'tailoring' derived from student interviews



	Count	Relevance
tailoring	82	100%
personalisation	32	39%
relevance	28	34%
specificity	22	27%
content	18	22%
adaptability	18	22%
feedback	12	15%
context	12	15%
understanding	11	13%
needs	11	13%
guidance	11	13%
support	9	11%

4.2.7 Instructor Remarks on AI Feedback

Nine instructors participating in the programme were recruited to provide comment on the AI feedback provided to students. They added linear notes to digital documents containing the AI feedback on student work, providing their own evaluations of the AI-generated feedback they received. Following its submission to *Leximancer* and the application of its algorithm, four main themes were arrived at based on their most prevalent comments (fig. 2).

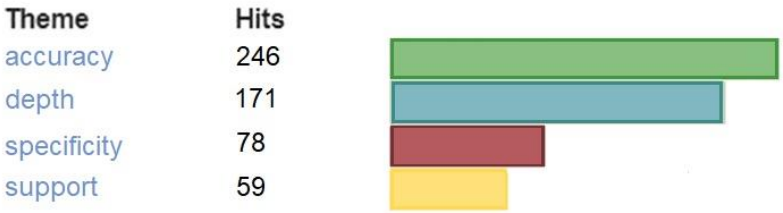
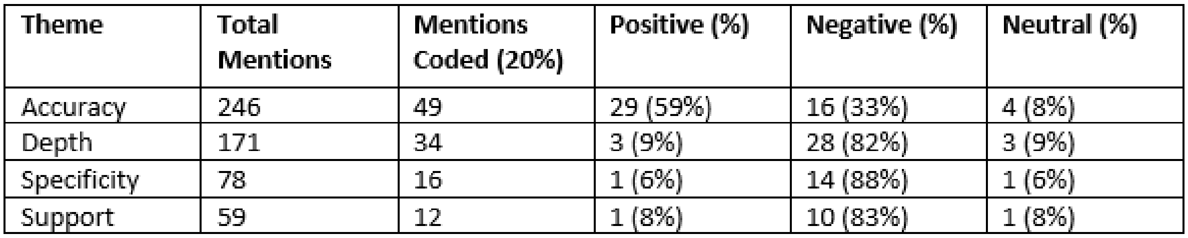


Figure 2: *M*ajor themes derived from instructor interviews with illustration of conceptual frequency

Table 12: Sentiment-based content analysis for main concepts/themes derived from instructor interviews



Theme	Total Mentions	Mentions Coded (20%)	Positive (%)	Negative (%)	Neutral (%)
Accuracy	246	49	29 (59%)	16 (33%)	4 (8%)
Depth	171	34	3 (9%)	28 (82%)	3 (9%)
Specificity	78	16	1 (6%)	14 (88%)	1 (6%)
Support	59	12	1 (8%)	10 (83%)	1 (8%)

Applying sentiment analysis to the four main themes revealed discrepancies between how instructors felt about the accuracy of AI grading versus the depth, specificity, and support offered by AI feedback (Table 12). Over half of those interviewed expressed positive sentiments with respect to accuracy, which, as is highlighted below, often referred to grading and specific recommendations for improvements. However, the vast majority of instructors also felt that the AI feedback lacked depth, was not specific enough, and did not sufficiently support students to make revisions. Details of their responses and reasoning are offered below.

4. 2.8 Accuracy

The instructor remarks on the AI feedback generated by the initiative indicates a broad acknowledgement that the feedback is often relevant. Numerous comments indicate that the feedback is 'on the right track', is 'good', or makes valid comment regarding the quality of student work. However, there are also criticisms that whilst the feedback is broadly accurate, it sometimes lacks nuance:

"The recommendation that the structure ought to be changed is valid but ultimately unnecessary." (Teacher 3 [T3])

"The comment is not entirely correct. Direct quotes can be a powerful way to support an argument, and it's not necessarily a problem to use them." (T2)

At times, the instructors would mark specific guidance as 'wrong' or 'incorrect', such as with respect to language and grammar. British English being corrected to American English was one example, as was the AI correcting 'learnt' to 'learned'.

It is notable also that instructors were able to see the AI mark assigned to student work alongside the AI feedback offered, as the two were provided in the same output when submitted to the AI programme. A number of instructor comments relating to 'accuracy' pertained to the marks or grades awarded as opposed to the accuracy of the *feedback* itself:

"I too would class this at around a 2:1. Though I wouldn't necessarily agree with every point made below… I think a 2:1 is fair." (T9)

It may be argued that if such remarks were excluded, the prevalence of 'accuracy' as a theme and its largely positive evaluation may be reduced in a repeat analysis. However, an holistic exploration of the comments tagged as related to 'accuracy' nevertheless does reveal a number of positive sentiments pertaining to feedback as compared with the three other main themes raised by instructors.

2.9 Depth

A related concern to remarks upon nuance was the concern that the feedback reviewed lacked 'depth' and that its engagement with the material was 'formulaic' and lacked in reference to specific examples:

"The comment on inadequate topic sentences is too general and does not provide specific examples from the student's paper to illustrate the point." (T1)

Across the comments pertaining to depth, instructors were overwhelmingly negative about the depth of the feedback offered by AI. There was a general sense across the comments left on AI feedback that AI responses were perfunctory and its judgments and recommendations superficial. One instructor accused it of having 'no real understanding' of the topic and another stated that it only 'scratch[ed] the surface' of what was wrong with one student essay. Instructors also expressed a lack of belief in the capacity of AI to analyse student work with sufficient depth when rendering judgment and providing feedback.

2.10 Specificity

AI was criticised in many places for its lack of specificity, offering generalised advice without direct reference to areas of the text. One instructor observed that AI did not appear to place comment on the original text but rather wrote a brief essay itself summarising the perceived problems with the essay. Several commentaries contained the observation that without making reference to specific areas within a student's work, it was not possible to offer useful insight into why a certain judgment or grade had been reached. One instructor argued that the feedback they had read 'could have been written about any essay' despite agreeing with the overall grade awarded.

As well as being inconsistent in terms of its depth of insight, there were concerns raised as well that the lack of specificity in recommendations for revisions also indicated a lack of pedagogical value. One remark stated this problem with respect to a specific student:

"The feedback has acknowledged that the student has not demonstrated sufficient critical evaluation, it makes recommendation only to the content of this rather than *how* to undertake critical evaluation. The student is exhibiting a lack of knowledge and/or skills that the feedback does little to address." (T2)

This evaluation suggests that AI lacks insight into student capabilities and skills, offering generalised advice as to how a person might improve a piece of work, but not as to how that *specific* person might improve their work. The impersonal nature of AI feedback was thus linked conceptually to a lack of specificity in its guidance on student revisions.

2.11 Support

As is hinted at above, instructors expressed a number of concerns about support for students how general AI advice proved in places:

"I cannot think that this is intended nor suitable for the replacement of the lecturer's guidance. The application obviously does not know the individual and cannot get to know them. It cannot identify what is apparent from this piece of work, which is that there are clear gaps in the student's linguistic ability that need to be addressed. This guidance would not be of use to the student alone as it cannot help them develop the skills needed to implement it." (T5)

The concern inferred from these remarks may be that instructors feel that AI feedback is not suitably tailored to student needs in order to help them develop their academic skills, giving an indication as to *what* they need to do it, but not *where* in the essay nor *how* to accomplish it.

The suggestion that AI cannot offer the same support for actually developing these skills as can instructors through their feedback highlights also the disconnection between the AI programme and instructors involved in the course, as AI recommendations were not passed onto instructors in order to alert them to specific errors made by students nor did it alert instructors to more general gaps in specific students' knowledge or ability. This highlights part of the function of instructor grading and feedback, which is to identify where students need support to develop further, continuing this support beyond the point of feedback itself. The implications of this for the design of AI grading and feedback initiatives may be considered in more depth below.

5. DISCUSSION

The findings from the qualitative analysis may be contrasted with that of the quantitative analysis. Both Spearman's correlation coefficient and the ICC tests suggest a high degree of accuracy for AI marking when compared with human marking, though this is not necessarily recognised in the interviews with students and instructor remarks on AI feedback. Across these findings, there is a general negativity about the prospect for generative AI to provide the same quality of marking or feedback as instructors. However, at least with regards to the accuracy of marking – something questioned by the students across the interviews but generally endorsed by instructors – this is not corroborated by the statistical analysis carried out as part of this study.

The issue of negative student and instructor attitudes towards AI in education may be addressed from the perspective of stakeholder education. As both student and instructor perspectives may be subject to assumptions or misunderstandings about how the AI initiative was intended to function and what were its ultimate goals were, they may have benefited from a clearer understanding of what the initiative was intended to achieve and as to what AI can and cannot provide. Providing onboarding sessions or instructional resources for either group may have contributed towards producing more trust in AI systems and fostering better informed evaluations of its processes with respect to institutional goals. This may be perceived as fulfilling a mandate for transparent communication, ensuring that stakeholders are kept abreast of initiative development, what it is intended to achieve, and what it entails for them as stakeholders. This might help stakeholders shift from a stance of resistance towards one of acceptance and collaboration.

5. 1. Participatory Evaluation

The discrepancies between the findings of the quantitative and qualitative research may be attributed in part to the research instruments used. The interviews with students and the inclusion of attitudes, experience and beliefs in the data provided have not only highlighted nuance in the students' perspectives but also have highlighted potential inadequacies with the initiative that the quantitative analysis has not discovered. Unfortunately, the focus on paper scores only in the quantitative side to the analysis has excluded statistical analysis as to the prevalence, strength and significance of the perspectives uncovered through interviews with students. Adding a further interrogatory mechanism – such as surveys of student perspectives suitable for statistical analysis – might better elaborate on the significance of the inadequacies raised through the qualitative analysis.

A further question is to what extent stakeholders were involved in the development of the initiative in a Hong Kong university. Participatory evaluation highlights the importance of including stakeholders in the design of initiatives and interventions, with participation correlated with engagement in the initiative/intervention (Suarez-Balcazar, 2003). Consulting students and instructors about how they might like to see AI used and integrated in future initiatives may succeed in creating a more positive assessment of its contributions. The findings of this study suggest that students view AI as having some role to play but as effectively offering inferior feedback and guidance as to revisions as compared to instructors and tutors. It may be anticipated that any initiative that was perceived as replacing instructor feedback as opposed to *complementing* instructor feedback might precipitate strong resistance.

This appears to be reflected in the concerns expressed by both instructors and students and reflected in the sentiment analysis of the responses associated with their respective main themes. Instructors expressed concerns about the depth and specificity of AI feedback as well as the lack of support that AI can offer to help students develop the skills required to make progress in their English assessments. The initiative was designed with instructors largely 'out of the loop' in the procedures of AI marking and feedback, meaning that they were detached from one of their usual avenues of assessing student ability and progress. The guidance presented to students by AI thus makes no reference to their past abilities and achievements, as well providing no avenue for support in implementing the recommended changes and closing skills gaps. An initiative that included a design that kept instructors within this loop might have received better evaluative responses from instructors, at least in terms of the support provided by AI feedback.

5.2. Theory of Change (ToC) Model

How to best design and implement such a programme may be explored from the perspective of ToC. This approach holds that it is important not only to understand stakeholder goals, but to understand the pathways and conditions that stakeholders perceive as necessary to attain these ends (Clark, 2004). From this perspective, if the goal is to reduce instructor workload whilst helping students revise their work effectively, it is clear that student dissatisfaction may prove to be one of the outcomes. The activities of the initiative may benefit from being more limited, such as being made available to students as a tool to use when undergoing draft revisions prior to seeking instructor feedback. On the other hand, the strong empirical evidence for AI marking as equivalent in accuracy to grading performed by instructors suggests its potential utility towards grading papers on instructors' behalf. However, it is unknown how instructors and students would react to this. It may be that instructor workload could be reduced by generative AI being tasked with performing secondary marking, reducing workload for instructors in classes that they do not teach or have little investment or involvement in.

From the perspective of ToC, it appears as though instructors may prefer to be kept 'in the loop' with regards to design and may resist their role being 'replaced' by AI – a preference that echoes student concerns about a lack of input or access to instructor feedback. Subsequent initiatives might see AI grading and reports relayed to instructors for instructors to consult and amend ahead of providing final grades and feedback. This effectively keeps instructors in the loop and may allay concerns regarding trust, tailoring, depth, and support. Such initiatives would benefit from stakeholder involvement in the planning and design phase as well as being subject to ongoing evaluation. It is for this reason that future initiatives may benefit from being designed and presented as ‘hybrid’ models rather than as attempts to reduce instructor workload by replacing the instructor in assessment roles perceived by both instructors and students as essential to the educational process.

5.3. RUFDATA Model

Future initiatives might also benefit from the application of RUFDATA in order to align the initiative with institutional goals, clarify the intended improvements offered by it, and to ensure that the relevant stakeholders are fully consulted in its development (Saunders, 2000). Following the rationale of iterative design, the initiative may be viewed as part of a cyclical process for designing an educational initiative that implements AI in assessment. The findings to this study ought not be interpreted as end-point criticisms but rather as data that can inform future redesigns as to the needs and expectations of users. Through employing RUFDATA in future iterative design, the institution can thereby embed stakeholder feedback within its ongoing monitoring and evaluation of the initiative, ensuring that it evolves over time in dialogue with the institution’s academic community. Future initiatives might also benefit from the application of RUFDATA in order to align the initiative with institutional goals, clarify the intended improvements offered by it, and to ensure that the relevant stakeholders are fully consulted in its development (Saunders, 2000). Following the rationale of iterative design, the initiative may be viewed as part of a cyclical process for designing an educational initiative that implements AI in assessment. The findings to this study ought not be interpreted as end-point criticisms but rather as data that can inform future redesigns as to the needs and expectations of users. Through employing RUFDATA in future iterative design, the institution can thereby embed stakeholder feedback within its ongoing monitoring and evaluation of the initiative, ensuring that it evolves over time in dialogue with the institution’s academic community.

5.4. Theoretical implications for design, implementation, and evaluation of AI-based educational assessment systems

The findings of this study offer critical insights into both the theoretical underpinnings and practical applications of AI-assisted grading and feedback systems in higher education.

5.4.1 Reconciling Technical Efficacy with Pedagogical Value

The study highlights a tension between the technical reliability of AI grading, evidenced by high inter-rater agreement with human markers and the perceived pedagogical limitations of AI feedback such as lack of depth, personalization, and contextual relevance. This aligns with cognitive load theory (Gonzaga et al., 2025) and assessment literacy frameworks (Zhan & Yan, 2025), which suggest that feedback efficacy depends on social cues, shared referents, and relational trust - elements inherently lacking in AI systems. All these imply future theories of AI in education must account for *dual metrics*: (1) objective accuracy (quantitative) and (2) subjective pedagogical utility (qualitative).

5.4.2 Stakeholder-Centric Design

The Theory of Change (ToC) and participatory evaluation frameworks underscore that AI systems must be co-designed with stakeholders (instructors and students) to align with institutional goals and user needs. This implies theoretical models should integrate RUFDATA (Saunders, 2000) to ensure evaluations address *Rationale,* Uses, Focus, Data, Audience, Timing, and Agency - bridging gaps between AI capabilities and stakeholder expectations.

5. 5. Practical implications for design, implementation, and evaluation of AI-based educational assessment systems

5.5.1 Hybrid Models for Feedback Delivery

Quantitative results support AI’s role in grading efficiency while qualitative data reveal strong stakeholder preference for human feedback’s nuance. It is recommended that "human-in-the-loop" models should be adopted where AI handles initial grading/routine feedback, in particular grammar and structure. Instructors should refine feedback for depth, context, and skill-building such as critical thinking and discipline-specific insights. For instance, AI could flag recurring issues like citation errors, while instructors provide tailored guidance on argumentation (Yun, 2025).

5.5.2 Stakeholder Education and Transparent Communication

Mistrust of AI stemmed partly from unfamiliarity. Students and instructors questioned AI’s reliability despite empirical evidence of grading accuracy. Recommendation is that pre-implementation workshops could be held in educational institutions to clarify AI’s role, limitations, and benefits such as timeliness and consistency. Transparent reporting of AI’s decision-making processes including rubric alignment and error rates is to be ensured to build trust between human and machines (Henderson et al., 2025).

5.6 Iterative Design and Participatory Evaluation

Addressing stakeholders' expressed concerns about AI's rigidity and lack of adaptability to individual needs, it is suggested that piloting AI tools in phased iterations, incorporating stakeholder feedback after each cycle like via surveys or focus groups is necessary. Another solution is to use participatory design to co-create feedback templates or customize AI outputs, for instance, allowing instructors to adjust AI-generated comments.

5.7 Enhancing AI Feedback Quality

As shown from the study, critiques of AI feedback's generality suggest a need for task-specific fine-tuning like discipline-aware LLMs and longitudinal adaptation such as AI systems that "learn" from instructor overrides or student revision patterns. For example, there could be the integration of self-assessment prompts (To, 2025) to encourage students to contextualize AI feedback.

5.8 Policy and Institutional Support

Lastly, successful implementation requires resource allocation for instructor training and AI system maintenance and ethical guidelines to address biases, data privacy, and accountability (Lindsay et al., 2025). The successful integration of AI-assisted grading and feedback systems hinges on two pillars: practical infrastructure and ethical governance. In terms of resource allocation, professional development programs must be provided for instructor training to equip them with the skills of interpreting, refining, and contextualizing AI-generated feedback. Such training should emphasize pedagogical integration, particularly aligning AI outputs with learning objectives. Another domain includes the ability to identify and correct AI errors or biases to overcome technical oversight of AI-based system. Institutions must invest in ongoing updates to AI models to address evolving curricular needs, language nuances, and disciplinary specificity for system maintenance.

In terms of ethical safeguards, regular audits of AI outputs for cultural, linguistic, or disciplinary biases (Lindsay et al., 2025), coupled with diverse training datasets, are critical to ensure equity. Transparent policies must govern student data usage, storage, and consent, adhering to regulations for data privacy. Clear protocols for challenging AI-generated grades or feedback - such as human review panels - can uphold academic integrity and stakeholder trust to solidify accountability frameworks. By addressing these operational and ethical challenges, institutions can foster AI systems that are not only efficient but also equitable and pedagogically meaningful. Future work should explore cost-benefit analyses of such implementations and their long-term impact on educational outcomes.

6 . CONCLUSION

This study has explored the utility of generative AI as part of a feedback system for correcting and marking academic work. Its objectives were: to ascertain how accurate AI-based grading is when compared with human feedback; to investigate how students experience and respond to AI-generated feedback; to explore how such systems might be more effectively implemented through the perspectives of ToC and participatory evaluation. Exploring an initiative implemented at a university in Hong Kong in Hong Kong, the study has examined both the accuracy of AI-generated essay marks as well as student experiences of AI generated feedback as part of an evaluation of the system informed by ToC and participatory evaluation.

The study reveals a high degree of inter-rater reliability between AI and human markers, suggesting the potential utility of AI in grading student papers. However, interviews with students suggest a conceptual focus on themes such as clarity, improvement, accuracy, trust and tailoring, with student perspectives across these themes being largely negative as to AI's performance during the initiative. Whilst some students appreciated the simplicity of the feedback, others felt it was sufficiently detailed or nuanced to support significant improvement in their drafts or progress in terms of their academic skills. Concerns about accuracy included factual errors and irrelevant feedback, whilst many expressed a lack of trust in AI, preferring human feedback due to its higher perceived degrees of accuracy and personally tailored feedback. Although students appreciated the iterative feedback from AI tools, they noted a lack of tailored insights. To enhance their learning experience, integrating specific self-assessment prompts, as suggested by To (2025), could enable students to evaluate source reliability and proofread drafts more effectively. Instructors also expressed concerns about the utility of AI to provide feedback on student work that holds sufficient depth, is specific in its criticisms, praise and recommendations, and that can support student development moving forwards.

The study demonstrates both the accuracy of AI in grading student papers as compared with human markers, but also significant scepticism and criticism among students and instructors alike as to its capacity to provide effective feedback on a par with instructors. Theories of change and participatory evaluation highlight the importance of liaising with stakeholders in system design, with this study highlighting the view that AI feedback should not be perceived as a 'replacement' for instructor feedback according to students. This implies its potential utility as a complementary component to a broader system integrating human and AI scoring and feedback. Ensuring that stakeholders are kept abreast of developments through stakeholder education and transparent communication may help encourage more positive sentiments surrounding future iterations of this initiative, which in turn should follow a structured approach to soliciting, monitoring and implementing stakeholder feedback into its design.

Beyond the need to include further stakeholder perspectives in such designs, there are a number of limitations with respect to the study's design. For one, the study did not measure and compare between the outcomes of feedback on revised work. The research above could be built upon by corroborating its interviews with statistical analysis of questionnaires posed to a larger cohort of students and/or instructors. Additionally, though sentiment analysis of leading themes was carried out with a sample of the data informing these themes, a full-scale sentiment analysis was beyond the scope of the study, meaning these findings are more indicative than exhaustive. Future research may benefit from sentiment analysis implemented at the point of thematic analysis to better evidence the evaluative stances of participants with respect to specific aspects of the initiative. Such improvements may inform future research design in line with a ToC approach emphasising the importance of participatory planning to successful educational initiative design, implementation and reception.

REFERENCES

Afzaal, M., Zia, A., Nouri, J. & Fors, U., 2024. Informative Feedback and Explainable AI-Based Recommendations to Support Students' Self-regulationTechnology, Knowledge and Learning. Technology, Knowledge and Learning, Volume 29, pp. 331-54. https://doi.org/10.1007/s10758-023-09650-0

Ahmed, I., Kajol, M., Hasan, U., Datta, P. P., Roy, A., & Reza, M. R. (2023). ChatGPT vs. Bard: a comparative study. UMBC Student Collection. https://doi.org/10.36227/techrxiv.23536290.v2

Allen, M., 2017. Survey: Open-Ended Questions. In: M. Allen, ed. The SAGE Encyclopedia of Communication Research Methods. London: Sage.

Atherton, P., Topham, L. & Khan, W., 2024. AI and student feedback. Edulearn 2024 - 16th International Conference on Education and New Learning Technologies, 1st Jul -3rd Jul 2024, Palma, Spain.

Banihashem, S. et al., 2024. Feedback sources in essay writing: Peer-generated or AI-generated feedback?. International Journal of Educational Technology in Higher Education, Volume 21. https://doi.org/10.1186/s41239-024-00455-4

Bauer, E. et al., 2023. Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda. British Journal of Educational Technology, 54(5), pp. 1222-45. https://doi.org/10.1111/bjet.13336

Bazeley, P., 2013. Qualitative Data Analysis: Practical Strategies. s.l.:Sage.

Bian, W., Alam, O. & Kienzle, J., 2020. Is automated grading of models effective?: Assessing automated grading of class diagrams. MODELS '20: Proceedings of the 23rd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems, pp. 365-76.

Bolderston, A., 2012. Conductive a Research Interview. Journal of Medical Imaging and Radiation Sciences, Volume 43, pp. 66-76. https://doi.org/10.1016/j.jmir.2011.12.002

Borade, J. & Netak, L., 2021. Automated Grading of Essays: A Review. Intelligent Human Computer Interaction, pp. 238-49. https://doi.org/10.1007/978-3-030-68449-5_25

Braun, V. & Clarke, V., 2006. Using thematic analysis in psychology. Qualitative Research in Psychology , 3(2), pp. 77-101.

Celik, I., Dindar, M., Muukkonen, H. & Jarvela, S., 2022. The Promises and Challenges of Artificial Intelligence for Teachers: A Systematic Review of Research. TechTrends 66, pp. 616-30. https://doi.org/10.1007/s11528-022-00715-y

Chan, S. T. S., Lo, N. P. K., & Wong, A. M. H. (2024). Enhancing university level English proficiency with generative AI: Empirical insights into automated feedback and learning outcomes. Contemporary Educational Technology, 16(4), Article ep541. https://doi.org/10.30935/cedtech/15607

Chang, D., Lin, M., Haian, S. & Wang, Q., 2023. Educational Design Principles of Using AI Chatbot That Supports Self-Regulated Learning in Education: Goal Setting, Feedback, and Personalisation. Sustainability, 15(17).

Chen, Z. & Wan, T., 2025. Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy. Phys. Rev. Phys. Educ., Volume 21.

https://doi.org/10.1103/PhysRevPhysEducRes.21.010126

Chu, Y. et al., 2025. A LLM-Powered Automatic Grading Framework with Human-Level Guidelines Optimisation. arXiv, 2410(02165).

Clark, H., 2004. Deciding the Scope of a Theory of Change. New York: ActKnowledge.

Crossley, S. et al., 2022. The persuasive essays for rating, selecting, and understanding argumentative and discourse element (PERSUADE) corpus 1.0. Assessing Writing 54.

Daniel, W., 1990. Spearman rank correlation coefficient. In: Applied Nonparametric Statistics. Ann Arbor: PW-Kent, pp. 358-65.

Dann, C. et al., 2024. Making sense of student feedback and engagement using artificial intelligence. Australasian Journal of Educational Technology, 40(3), pp. 58-76. https://doi.org/10.14742/ajet.8903

Davies, R., 2018. Representing theories of change: Technical challenges with evaluation consequences. Journal of Development Effectiveness 10(4), pp. 438-61.

Dimari, A. et al., 2024. AI-Based Automated Grading Systems for open book examination system: Implications for Assessment in Higher Education. 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS).

Evans, C., 2018. Analysing Semi-Structured Interviews Using Thematic Analysis. London: Sage.

Fleckenstein, J., Liebenow, L. & Meyer, J., 2023. Automated feedback and writing: A multi-level meta-analysis of effects on students’ performance. Frontiers in Artificial Intelligence 6. https://doi.org/10.3389/frai.2023.1162454

Gao, R. et al., 2024. Automatic assessment of text-based responses in post-secondary education:. Computers and Education: Artificial Intelligence 6. https://doi.org/10.1016/j.caeai.2024.100206

Geigle, C., Zhai, C. & Ferguson, D., 2016. An Exploration of Automated Grading of Complex Assignments. L@S '16: Proceedings of the Third (2016) ACM Conference on Learning @ Scale, pp. 351-60.

Gergen, K., 2015. An Invitation to Social Construction. 3rd ed. LA: Sage.

Gnepp, J., Klayman, J., Williamson, I. & Barlas, S., 2020. The future of feedback: Motivating performance improvement through future-focused feedback. PLoS One 15(6). https://doi.org/10.1371/journal.pone.0234444

Golchin, S., Garuda, N., Impey, C. & Wenger, M., 2024. Grading Massive Open Online Courses Using Large Language Models. arXiv, 2406(11102).

Gonzaga, J., Jiang, Y. & Vassar, A., 2025. Empowering CS1 Educators: Enhancing Automated Feedback Instruction with Cognitive Load Theory. SIGCSETS 2025: Proceedings of the 56th ACM Technical Symposium on Computer Science Education, Volume 2, pp. 1461-2.

Graham, S., Hebert, M. & Harris, K., 2015. Formative assessment and writing. The Elementary School Journal 115(4), pp. 523-47.

Gratani, F. et al., 2024. Personalised Feedback in University Contexts: Exploring the Potential of AI-Based Techniques. Higher Education Learning Methodologies and Technologies Online, pp. 440-54.

Grevisse, C., 2024. LLM-based automatic short answer grading in undergraduate medical education. BMC Medical Education, 24(1060). https://doi.org/10.1186/s12909-024-06026-5

Guo, K., Zhang, E., Li, D. & Yu, S., 2025. Using AI-supported peer review to enhance feedback literacy: An investigation of students' revision of feedback on peers' essays. British Journal of Educational Technology, 56(4), pp. 1612-39. https://doi.org/10.1111/bjet.13540

Hahn, M., Navarro, S., La Fuente Valentin, I. & Burgos, D., 2021. A systematic review of the effects of automatic scoring and automatic feedback in educational settings. IEEE Access, p. 9.

Hammersley, M., 2015. Sampling and thematic analysis: a response to Fugard and Potts. International Journal of Social Research Methodology 18(6), pp. 687-8. https://doi.org/10.1080/13645579.2015.1005456

Heaven, W., 2022. Generative AI is changing everything. But what’s left when the hype is gone?. [Online]
Available at: https://www.technologyreview.com/2022/12/16/1065005/generative-ai-revolution-art/
[Accessed 16 June 2025].

Henderson, M. et al., 2025. Comparing Generative AI and teacher feedback: Student perceptions of usefulness and trustworthiness. Assessment & Evaluation in Higher

Education . https://doi.org/10.1080/02602938.2025.2502582

Holmes, A., 2020. Researcher Positionality -- A Consideration of Its Influence and Place in Qualitative Research -- A New Researcher Guide. Shanlax International Journal of Education, 8(4), pp. 1-10. https://doi.org/10.34293/

Hooda, M., Rana, C., Dahiya, O., Rizwan, A., & Hossain, M. S., 2022. Artificial intelligence for assessment and feedback to enhance student success in higher education. Mathematical Problems in Engineering, 2022(1), 5215722.

Impey, C. et al., 2025. Using Large Language Models for Automated Grading of Student Writing about Science. International Journal of Artificial Intelligence in Education. https://doi.org/10.1007/s40593-024-00453-7

Ivankova, N., Creswell, J. & Stick, S., 2006. Using Mixed-Methods Sequential Explanatory Design: From Theory to Practice. Field Methods.

Jung, H., 2019. The Evolution of Social Constructivism in Political Science: Past to Present. SAFE Open, 9(1).

Knoth, N., Tolzin, A., Janson, A. & Leimeister, J., 2024. AI literacy and its implications for prompt engineering strategies. Computers and Education: Artificial Intelligence 6. https://doi.org/10.1016/j.caeai.2024.100225

Koch, G., 1982. Intraclass correlation coefficient. In: Encyclopedia of Statistical Sciences, Vol. 4. New York: John Wiley, p. 213–217.

Kritt, D., 2018. Teaching As if Children Matter. In: D. Kritt, ed. Constructivist Education in an Age of Accountability. London: Palgrave Macmillan, pp. 3-19.

Lagakis, P., Demetriadis, S. & Psathas, G., 2024. Automated Grading in Coding Exercises Using Large Language Models. Smart Mobile Communication & Artificial Intelligence, pp. 363-73. https://doi.org/10.1007/978-3-031-54327-2_37

Lee, A., 2023. Supporting students' generation of feedback in large-scale online course with artificial intelligence-enabled evaluation. Studies in Educational Evaluation, Volume 77.

Lee, D. et al., 2024. The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. Computers and Education: Artificial Intelligence 6. https://doi.org/10.1016/j.caeai.2024.100221

Lee, J. & Song, Y., 2024. College Exam Grader using LLM AI models. 2024 IEEE/ACIS 27th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD).

Li, H. et al., 2025. LLM-based Automated Grading with Human-in-the-Loop. arXiv, 2504(05239).

Liljequist, D., Elfving, B. & Roaldsen, K., 2019. Intraclass correlation – A discussion and demonstration of basic features. PLoS One, 14(7).

Lindsay, E., Zhang, M., Johri, A. & Bjerva, J., 2025. The Responsible Development of Automated Student Feedback with Generative AI. 2025 IEEE Global Engineering Education Conference (EDUCON).

Lo, N., Wong, A., and Chan, S. (2025). The impact of generative AI on essay revisions and student engagement. Computers and Education Open 22:100249. doi: 10.1016/j.caeo.2025.100249

Machado, A. et al., 2025. Workload perception in educational resource recommendation supported by artificial intelligence: A controlled experiment with teachers. Smart Learning Environments, Volume 12. ttps://doi.org/10.1186/s40561-025-00373-6

Madigan, D. & Kim, L., 2021. Does teacher burnout affect students? A systematic review of its association with academic achievement and student-reported outcomes. International Journal of Educational Research 105. https://doi.org/10.1016/j.ijer.2020.101714

Makgamatha, M., 2009. Challenges in implementing a participatory evaluation approach: A case study of the Limpopo Literacy Teaching Evaluation Project. Education as Change 13(1), pp. 91-103. https://doi.org/10.1080/16823200902940730

Mertens, U., Finn, B. & Lindner, M., 2022. Effects of computer-based feedback on lower- and higher-order learning outcomes: A network meta-analysis. Journal of Educational Psychology 114(8), pp. 1743-72. https://doi.org/10.1037/edu0000764

Messer, M., Brown, N., Kolling, M. & Shi, M., 2023. Machine Learning-Based Automated Grading and Feedback Tools for Programming: A Meta-Analysis. ITiCSE 2023: Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education , pp. 491-7.

Meyer, J. et al., 2024. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. Computers and Education: Artificial Intelligence, Volume 6. https://doi.org/10.1016/j.caeai.2023.100199

Mourby, M. et al., 2019. Governance of academic research data under the GDPR - lessons from the UK. International Data Privacy Law, 9(3), pp. 192-206.

Nazaretsky, T. et al., 2024. AI or Human? Evaluating Student Feedback Perceptions in Higher Education. OSF Preprints.

Nazaretsky, T. et al., 2024. AI or Human? Evaluating Student Feedback Perceptions in Higher Education. Technology Enhanced Learning for Inclusive and Equitable Quality Education, pp. 284-98. https://doi.org/10.1007/978-3-031-72315-5_20

Noble, L., Wong, A., & Chan, S. (2025). The impact of generative AI on essay revisions and student engagement. Computers and Education Open, 100249. https://doi.org/10.1016/j.caeo.2025.100249

Nowell, L., Norris, J. & White, D., 2017. Thematic Analysis: Striving to Meet the Trustworthiness Criteria. International Journal of Qualitative Methods, 16(1), pp. 1-13.

Oliver, P., 2010. The Student's Guide to Research Ethics. Milton Keynes: Open University Press.

Palinkas, L. et al., 2016. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health. , 42(5), pp. 533-44.

Petersen, T., 2024. Students prefer teacher feedback over AI feedback, research finds. [Online]
Available at: https://phys.org/news/2024-09-students-teacher-feedback-ai.html

Peters, K. & Halcomb, E., 2015. Interviews in qualitative research. Nurse Researcher, 22(4), pp. 6-7. https://doi.org/10.7748/nr.22.4.6.s2

Pulla, V. & Carter, E., 2018. Employing Interpretivism in Social Work Research. International Journal of Social Work and Human Services Practice, 6(1), pp. 9-14.

Qin, D., 2016. Positionality. In: The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies. London: John Wiley & Sons.

Reinholz, D. & Andrews, T., 2020. Change theory and theory of change: What's the difference anyway?. International Journal of STEM Education 7(2). https://doi.org/10.1186/s40594-020-0202-3

Resnik, D., 2020. What is ethics in research & why is it important?. [Online]
Available at: https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm

Ruiz-Palmero, J., Fernández-Lacorte, J., Sánchez-Rivas, E. & Colomo-Magaña, E., 2020. The implementation of Small Private Online Courses (SPOC) as a new approach to education. International Journal of Educational Technology in Higher Education 17(27). https://doi.org/10.1186/s41239-020-00206-1

Sai, S., Kanadia, M. & Chamola, V., 2024. Empowering IoT with Generative AI: Applications, Case Studies, and Limitations. IEEE Internet of Things Magazine 7(3). https://doi.org/10.1109/IOTM.001.2300246

Salcedo, J. & McCormick, K., 2020. SPSS Statistics. 4th ed. Hoboken: John Wiley.

Saunders, B., Kitzinger, J. & Kitzinger, C., 2015. Anonymising interview data: Challenges and compromise in practice. Qualitative Research, 15(5), pp. 616-21.

Saunders, M., 2000. Beginning an Evaluation with RUFDATA: Theorising a Practical Approach to Evaluation Planning. Evaluation 6(1), pp. 7-21. https://doi.org/10.1177/1468794114550439

Schneider, J., Schenk, B. & Niklaus, C., 2024. Towards LLM-based Autograding for Short Textual Answers. Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024), 2309(11508).

Shi, H. & Aryadoust, V., 2024. A systematic review of AI-based automated written feedback research. ReCALL, 36(2), pp. 187-209. https://doi.org/10.1017/S0958344023000265

Smith, A. & Humphreys, M., 2006. Evaluation of unsupervised semantic mapping of natural. Behaviour Research Methods 38(2), pp. 262-79. https://doi.org/10.3758/BF03192778

Spearman, C., 1904. The Proof and Measurement of Association between Two Things. The American Journal of Psychology 15(1), pp. 72-101.

Suarez-Balcazar, Y., 2003. Empowerment and Participatory Evaluation of Community Interventions: Multiple Benefits. New York: Hawarth Press.

Su, N., 2018. Positivist Qualitative Methods. In: C. Cassell, A. Cunliffe & G. Grandy, eds. The SAGE Handbook of Qualitative Business and Management Research Methods: History and Traditions. London: Sage.

Tishenina, M., 2024. The broken pillar: AI for feedback generation and the erosion of students' trust. [Online] Available at: https://www.bera.ac.uk/blog/the-broken-pillar-ai-for-feedback-generation-and-the-erosion-of-students-trust

To, K. H., 2025, June 17. Three ways to get students using GenAI in project-based learning. THE - Times Higher Education. https://www.timeshighereducation.com/campus/three-ways-get-students-using-genai-projectbased-learning

Vetrivel, S., Arun, V., Ambikapathi, R. & Saravanan, T., 2025. Automated Grading Systems: Enhancing Efficiency and Consistency in Student Assessments. In: T. Murugan, K. Periasamy & A. Abirami, eds. Adopting Artificial Intelligence Tools in Higher Education. New York: CRC Press.

Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. Expert Systems with Applications, 252, 124167. https://doi.org/10.1016/j.eswa.2024.124167

Witkin, S., 2012. An Introduction to Social Constructions. In: S. Witkin, ed. Social Construction and Social Work Practice: Interpretations and Innovations. New York: Columbia University Press, pp. 13-37.

Xiao, C. et al., 2025. Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs. LAK '25: Proceedings of the 15th International Learning Analytics and Knowledge Conference, pp. 293-305.

Xie, W., Niu, J., Xue, C. & Guan, N., 2024. Grade Like a Human: Rethinking Automated Assessment with Large Language Models. arVix, 2405(19694).

Xu, W. et al., 2023. Artificial intelligence in constructing personalised and accurate feedback systems for students. International Journal of Modeling, Simulation, and Scientific Computing, 14(1). https://doi.org/10.1142/S1793962323410015

Yeung, C. et al., 2025. A Zero-Shot LLM Framework for Automatic Assignment Grading in Higher Education. arXiv, 2501(14305).

Yun, C. S., 2025. Bridging AI and Human Expertise in ESP: Optimizing AI-Integrated Instruction for Job Application Skills. ESP Review, 7(1), 53-77. https://doi.org/10.23191/espkor.2025.7.1.53

Zhang, Y. & Wildemuth, B., 2017. Unstructured Interviews. In: B. Wildemuth, ed. Applications of Social Research Methods to Questions in Information and Library Science. Santa Barbara: Libraries Unlimited, pp. 239-47.

Zhan, Y. & Yan, Z., 2025. Students' engagement with ChatGPT feedback: Implications for student feedback literacy in the context of generative artificial intelligence. Assessment & Evaluation in Higher Education . https://doi.org/10.1080/02602938.2025.2471821
한수미 and 김민지. (2025). 생성형 AI 기반 영어학개론 수업 설계를 위한 기초 연구. ESP Review, 7(1), 99-118. https://doi.org/10.23191/espkor.2025.7.1.99

APPENDIX A

Interview questions for students

1. How comprehensible do you find the AI generated feedback as compared with human instructors?
2. In what ways has the AI generated feedback helped you improve your work?
3. How tailored do you find the AI generated feedback is with respect to the content of your work?
4. How do you feel about the accuracy of the AI generated feedback with respect to identifying areas for improvement within your work?
5. To what extent do you trust the AI generated feedback as compared with feedback from human instructors?
6. How has the AI generated feedback influence your motivation and engagement with the course?
7. In what ways do you feel the AI generated feedback could be improved?
8. How do you feel about AI generated gradings being used to grade written work?
9. How easy or difficult is it to implement the suggestions given to you by AI generated feedback?
10. Would you prefer to receive feedback solely from AI, from human instructors, or from a combination of both, and why?

APPENDIX 2

Interview questions for instructors

1. How accurate do you find the AI-generated grading compared to your own assessments?

2. How effective do you find the AI-generated feedback in helping students improve their work?
3. Do you feel the AI-generated feedback provides sufficient depth and clarity for students?
4. How well does the AI-generated feedback align with the course rubrics and marking criteria?
5. To what extent do you believe AI-generated feedback is tailored to individual students' needs?
6. How do you feel about the role of AI in reducing instructor workload?
7. What are the limitations of AI feedback, based on your observations?
8. How do you think AI-generated feedback could be improved for better alignment with educational goals?
9. How do you feel about the potential for AI to replace or supplement human feedback in assessments?
10. Do you believe students value AI feedback as much as they value feedback from human instructors? Why or why not?