Computers and Education: Artificial Intelligence 10 (2026) 100511

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

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

Stretching AI's reach: Assessing an AI-driven feedback system for extended academic writing

Jim Lo [a,*], Christy Wong [a], Agnes Ng [a], Pinna Wong [a], Denise Cheung [a], Pauli Lai [b]

[a] English Language Centre, The Hong Kong Polytechnic University, Hong Kong, China
[b] Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

Keywords:
Automated writing evaluation
AI-Generated feedback
Academic writing
Rubric-based feedback
Hybrid intelligence

ABSTRACT

Advances in large language models (LLMs) enable timely and scalable writing evaluation. Previous research has shown that LLM-driven conversational systems, such as ChatGPT, can provide feedback on short essays. However, it is unclear whether AI can effectively evaluate more demanding genres. This study investigates a custom-built writing feedback system developed at a Hong Kong university that uses OpenAI's GPT-4 Turbo (0125-preview) to provide rubric-based feedback on a 1500-word academic report. Guided by a detailed, rubric-aligned prompt, the system generated 333 feedback items from 37 undergraduates, which were analysed for accuracy, tone, and inclusion of examples. The analysis showed that most feedback was accurate and addressed both strengths and weaknesses, but over half lacked concrete examples. Often recycling phrases from rubric descriptors, the feedback was largely generic and occasionally inaccurate. Interview data from six students revealed that the AI feedback was valued for its coverage, efficiency, and constructive tone, yet its generic nature undermined its usefulness. Despite these limitations, students expressed interest in receiving both AI and teacher feedback for the efficiency and coverage that AI offers, alongside the specificity and relevance of teacher input. These findings suggest that employing a well-crafted prompt on an AI model with a large context window does not necessarily guarantee substantive feedback. Therefore, educators using AI-driven feedback systems should thoroughly assess these systems' capacity to handle extended academic writing. Future research could explore ways to refine prompts and system design for long-form writing assignments.

1. Introduction

Large language models (LLMs), most notably OpenAI's ChatGPT and its various successors, have transformed natural language processing. They can reason, summarise, and generate coherent responses to user input (Bubeck et al., 2023). These capabilities hold considerable promise for writing evaluation. Underpinning these capabilities is LLMs' transformer architecture, which processes text in parallel and predicts the next word to generate contextually relevant responses (Vaswani et al., 2017). Unlike sequential processing, parallel processing enables LLM-based generative AI systems to analyse extended texts swiftly and provide timely, multi-stage feedback at scale, thereby strengthening student engagement and promoting continuous learning. Inspired by these affordances, a growing body of research now investigates whether AI can provide feedback effectively (e.g., Banihashem et al., 2024; Dai et al., 2024; Escalante et al., 2023; Guo & Wang, 2024; Lin & Crosthwaite, 2024; Steiss et al., 2024; Wang, 2024). These studies, however, have largely focused on short essays, leaving open questions about how well AI performs on extended academic writing. Long-form academic writing poses qualitatively different demands from short essays. Beyond grammatical accuracy and paragraph cohesion, students must manage rhetorical moves across sections, integrate sources, and sustain coherence over long portions of text (Hyland, 2008; Swales, 1990). At the system level, processing texts longer than those on which LLMs were trained can be challenging due to architectural constraints that make it difficult to consistently access information from the entire input (Liu et al., 2024; Wang et al., 2024). Therefore, findings from studies on short essays may not generalise to longer academic texts. To address this gap, the present study evaluates a custom-built AI feedback system with a rigorously designed prompt applied to a longer academic report. By examining the system's performance in delivering accurate, specific, and comprehensive feedback, this study offers empirical insight into the potential and limitations of AI-generated feedback for extended academic writing, as well as practical design guidance for implementing

* Corresponding author. Room AG612, Core A, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China.
E-mail address: jim.lo@polyu.edu.hk (J. Lo).

https://doi.org/10.1016/j.caeai.2025.100511
Received 6 May 2025; Received in revised form 4 November 2025; Accepted 25 November 2025
Available online 28 November 2025

a custom-built AI feedback system.

2. Literature review

2.1. AI in language education

The capabilities of generative AI have opened new possibilities in language education. Because large language models (LLMs) are designed for natural language interaction, students can communicate with them in everyday language. This feature can encourage autonomous learning by enabling students to create their own content (Lee et al., 2023) and develop effective self-study routines outside the classroom (Van Horn, 2024). In addition to text-based interaction, AI-powered conversational tools with speech recognition and text-to-speech technology allow users to speak to a system and receive synthesised spoken responses. These tools not only support listening and speaking practice in academic settings (Kohnke et al., 2025) but also create low-stress environments where students can make mistakes and receive feedback without fear of negative judgement (Crompton et al., 2024).

Alongside its role in facilitating autonomous learning, generative AI is proving to be a powerful ally for teachers in creating and personalising teaching materials. Trained on vast amounts of text, LLMs can rapidly produce texts tailored to genre, topic, and style (Moorhouse, 2024). Customising lesson content empowers teachers to address students' interests and learning needs, which is especially beneficial for differentiating instruction in heterogeneous classes (Kohnke & Zou, 2025). Beyond saving time on producing materials, AI's rapid responses make it possible to assess students adaptively in real time, with question difficulty dynamically adjusted based on student performance. This adaptive questioning could gauge students' comprehension abilities more fairly by preventing them from being under-challenged or overwhelmed (Khlaif et al., 2024). Therefore, although concerns remain about possible effects on students' critical thinking (Ma et al., 2024) and academic integrity (Chan, 2024), generative AI is already widely used in language education and warrants ongoing study (Wang et al., 2025).

2.2. Feedback on writing

Feedback is generally recognised as a crucial element in developing students' writing and consolidating their learning. It guides students in bridging the gap between their current abilities and higher standards, while introducing them to the conventions and expectations of disciplinary genres (Hyland & Hyland, 2019). Feedback also helps students plan improvements and monitor how they use learning strategies across different tasks (Evans, 2013). Research has identified effective feedback practices that can deepen student engagement, such as timeliness (Bayerlein, 2014) and multi-stage feedback (Ferris, 2022). Timeliness refers to delivering feedback when the writing is still salient to students, which can motivate them to implement the feedback (Fisher et al., 2025). Multi-stage feedback, delivered through cycles of drafting, feedback, and revision, encourages students to address different aspects of their writing in each cycle (Zhang & Hyland, 2022). Effective feedback should also balance coverage of global and local issues rather than focusing predominantly on local ones (Lee, 2008), and it should highlight strengths as well as problems so that students can be empowered to apply successful strategies in future tasks (Goldstein, 2004). However, providing broad feedback that includes both praise and constructive criticism in a timely and iterative manner is labour-intensive, and large class sizes and heavy teacher workloads make delivering such feedback particularly challenging (Ferris & Hedgcock, 2014). AI offers a promising solution to these challenges by enabling rapid and personalised feedback at scale. This has prompted extensive research to explore the feasibility of implementing AI-driven feedback (e.g., Banihashem et al., 2024; Dai et al., 2024; Escalante et al., 2023; Guo & Wang, 2024; Lin & Crosthwaite, 2024; Steiss et al., 2024; Venter et al., 2025; Wang, 2024).

2.3. AI feedback on writing

Research has demonstrated the strengths of AI-generated feedback on writing. AI feedback was found to be more detailed and consistently more readable than human instructor feedback (Dai et al., 2024). Among lower-intermediate English learners, AI feedback was perceived as accurate and useful, particularly for local aspects such as style and vocabulary (Escalante et al., 2023). Beyond language concerns, broad coverage appeared to be an advantage of AI feedback. Compared with teachers' comments, AI feedback distributed attention more evenly across content, organisation, and language (Guo & Wang, 2024). It also provided more positive remarks acknowledging learners' progress, thereby inspiring them to act on the feedback in subsequent revisions (Banihashem et al., 2024; Guo & Wang, 2024). Being efficient and constantly available, AI encouraged learners to continuously improve their writing through multiple rounds of feedback (Wang, 2024). These findings suggest that AI can support effective feedback practices including multi-stage revision cycles and balanced attention to global and local issues, areas where teachers often struggle to provide feedback consistently.

However, these strengths may not ensure that the feedback is valid or useful for revision. A notable limitation is AI's tendency to prioritise local over global issues. Lin and Crosthwaite (2024) compared corrective feedback by teachers and GPT-4 on essays by upper-intermediate learners of English. Contrary to Guo and Wang (2024), they found that GPT-4 placed a disproportionate focus on local issues, often unnecessarily rewriting sentences that were already correct, causing over-correction. Targeting coherence and cohesion in essays by English learners at various levels, Yoon et al. (2023) similarly found that AI feedback tended to overemphasise surface-level features, such as phrase repetition, at the expense of deeper concerns about idea development. They also observed that AI often relied on superficial cues: the mere presence of connectives would lead it to judge a text as coherent despite obvious gaps in reasoning, whereas a coherent text without overt connectives was often flagged for insufficient use of transitional words. Such findings raise questions about whether AI can reliably provide valid feedback beyond sentence-level matters.

Another recurring concern is that AI feedback can be generic. Across various research contexts, AI has been found to provide formulaic suggestions rather than precise critiques, which limits the usefulness of its feedback for revision (Awidi, 2024; Su et al., 2023; Venter et al., 2025). While training LLMs on broad, non-domain-specific datasets may lead to generic feedback (Venter et al., 2025), imprecise prompts and rubrics can also contribute to it (Su et al., 2023). Furthermore, studies have documented cases where AI gave inaccurate feedback due to hallucinations. Hallucinations occur when AI generates plausible but incorrect content that is not grounded in the prompt or external sources (Bubeck et al., 2023; OpenAI, 2025). This has resulted in AI feedback asserting that a key point was missing when it was already present (Steiss et al., 2024) or reporting clear paragraphing when there was none (Pack et al., 2024). These issues highlight the need for careful prompt design, particularly by requiring AI to substantiate its feedback with examples from the text. Including text-based examples can turn generic comments into actionable revision suggestions. Examples also make mismatches between a claim and its cited evidence easy to identify, helping students disregard fluent but inaccurate feedback. Given these limitations, researchers recommend a prudent approach by using AI on early drafts (Steiss et al., 2024) and in low-stakes assessments (Tate et al., 2024).

Variations in research focus and prompt design further complicate the findings. Some studies targeted argumentation (Banihashem et al., 2024) or coherence and cohesion (Yoon et al., 2023). Others assessed AI feedback more generally, considering both global and local issues (Escalante et al., 2023; Gozali et al., 2024; Guo & Wang, 2024). Prompts ranged from simple requests to assess content and language (Guo & Wang, 2024; Lin & Crosthwaite, 2024) to elaborate instructions with task guidelines and rubric descriptors (Awidi, 2024). Some studies also assigned AI a role, for instance, a language assessment expert (Pack et al., 2024), a writing tutor (Su et al., 2023), or a secondary school teacher (Steiss et al., 2024). Consistent across these studies is that they mostly employed ChatGPT's conversational interface for feedback on short essays that were under 500 words. In addition, many studies did not include a rubric (e.g., Banihashem et al., 2024; Escalante et al., 2023; Guo & Wang, 2024; Lin & Crosthwaite, 2024; Steiss et al., 2024), which may have led to an excessive focus on particular aspects of writing (e.g., Lin & Crosthwaite, 2024). It is therefore uncertain whether the breadth and usefulness of AI feedback observed on short essays can generalise to longer writing, and whether a custom-built AI feedback system guided by a rubric could overcome some of the documented limitations.

2.4. Impetus for the present study

This study aims to investigate the feasibility of using a custom-built AI feedback system to provide feedback on extended academic writing. Having a specifically designed platform should reduce inconsistencies in feedback caused by ChatGPT's conversational interface (Lin & Crosthwaite, 2024). This design also circumvents the need for prompt engineering as teachers are often inexperienced in creating prompts (Ma et al., 2024; Moorhouse et al., 2024). To address the excessive focus on local issues, the system uses a prompt that includes a comprehensive nine-item rubric to ensure that AI balances its attention on all aspects of the writing. To counter generic feedback stemming from imprecise prompts and rubrics (Su et al., 2023), the prompt includes the assignment context and requires AI to support its feedback with examples from the text on each rubric item. Although requiring AI to ground its feedback in the text does not necessarily prevent hallucinations (Bubeck et al., 2023), it could make them easier to detect if the examples contradict the assertion of the feedback. In addition, there have been observations that AI tends to provide positive feedback even without explicit prompting (e.g., Banihashem et al., 2024; Guo & Wang, 2024). Yet, there seems to be little prior research that systematically compares the frequency of positive versus negative AI feedback. While positive remarks can enhance student motivation, little is known about whether AI feedback is overly positive. Because LLMs are often trained to generate supportive responses (Bai et al., 2022; Ouyang et al., 2022; Touvron et al., 2023), AI may tend to emphasise praise rather than critiques that highlight problems. This positivity bias could reduce the usefulness of feedback and warrants investigation.

Unlike most previous studies that evaluated AI feedback on short essays that were under 500 words (e.g., Awidi, 2024; Escalante et al., 2023; Guo & Wang, 2024; Lin & Crosthwaite, 2024; Yoon et al., 2023), this research examines academic report drafts of about 1500 words. Long-form academic writing is chosen because it places greater demands on rhetorical control, sustained argumentation, and source integration. Compared with short essays that often rely on formulaic paragraph structures and limited exposition, extended academic writing requires flexible use of multiple rhetorical moves across longer stretches of text to meet disciplinary expectations (Hyland, 2008; Swales, 1990). Whereas integration of evidence is often optional in short essays, it is essential in academic writing to synthesise previous research and contextualise the topic (Swales & Feak, 2012). Longer inputs also pose challenges for AI feedback: processing texts longer than those on which LLMs were trained may cause them to have difficulty reliably accessing information across the entire text input (Liu et al., 2024; Wang et al., 2024). Therefore, it remains unclear whether benefits observed on short essays can be transferred to extended academic writing. By addressing some of the common concerns regarding AI-generated feedback, this study aims to answer the following research question:

To what extent can a custom-built AI feedback system, guided by a detailed prompt, provide accurate, specific, and comprehensive feedback on extended academic reports?

3. Method

3.1. Research context and participants

The study was conducted within the context of a general education course at a university in Hong Kong. As part of the course requirements, students were asked to write a research report on a topic related to the history of Hong Kong. The report was based on a site visit, during which students evaluated how historical information was presented at the site. The assessment of the report was a collaborative effort between the course lecturer and the language teachers at the university's English language centre. Before submitting their final reports to the course lecturer, students received feedback on their report drafts from the AI-driven feedback system and the language teachers. The required length of the report draft was about 1500 words.

Thirty-seven undergraduate students from various disciplines participated in the study. With the university's subscription to multiple GenAI tools (e.g., GPT-4, Llama, and Copilot with Bing) for student use since 2023, they were likely to have some experience in using AI for feedback. Participation in the study was voluntary, and informed consent was obtained from all participants. The consent ensured that their report drafts and the associated AI feedback could be utilised for research purposes. To maintain anonymity, each student was assigned a unique identifier.

3.2. Instrument

The AI feedback system was developed by an academic department of the university. Given the length of the report draft and the detailed prompt, the system used GPT-4 Turbo (0125-preview) as its AI model. The model had a sufficiently large context window to process the draft and the prompt and generate detailed feedback.[1] The model's capacity should also enable the system to follow the prompt's instruction to provide specific examples from the draft to explain its strengths and weaknesses. The target users of the system were course lecturers across the university, so the interface had a user-friendly design without the need for prompt engineering. The main inputs from the user were the assignment context and the rubric, which would become part of the prompt. The prompt was sent to GPT-4 Turbo through code written in the programming language Python. In the code, the AI model was given the following information:

- the role of a helpful and outstanding university professor who provides objective and constructive feedback that encourages students to improve their writing
- the assignment context summarising the writing task and expectations
- the nine-item rubric with top-level descriptors
- the submitted draft

With the above information, it was instructed to perform the following tasks:

- give feedback and assign a grade ("A+" to "F") for each rubric item, and assign an overall grade
- include specific examples from the draft to explain strengths and weaknesses
- evaluate the draft rigorously and base feedback on the rubric
- assess the draft's relevance to the assignment context

[1] GPT-4 Turbo allows up to 128,000 tokens for input and 4096 for output (OpenAI, 2024). Assuming an average of 0.75 words per token, these token limits translate to roughly 96,000 words for input and 3072 words for output, which is more than enough for the purposes of the study.

- alert students to align their writing with the expectations if the draft deviates from the assignment context

The rubric included only top-level descriptors because the system was found to differentiate performance levels more effectively with these alone than when using descriptors for all levels. Since the assignment context and the rubric were customisable and each submission was different, these were treated as variables in the code. This resulted in prompts that were dynamically generated from these variables. The AI model was instructed to return its responses in JSON format for the system to process before displaying the grades and feedback. The temperature, a parameter that controls how random the model's responses are, was set to zero to minimise randomness and produce consistent feedback (Xu et al., 2024). Given the diversity of assignment topics and the simultaneous use of the system in various courses, the system adopted zero-shot prompting (Brown et al., 2020). This means that the model received the assignment context, rubric, and submitted draft, but no examples of student drafts or feedback were included as demonstrations. The model generated feedback by drawing upon its extensive pretraining to interpret the provided context and rubric in relation to the draft. This setup approximates common classroom use where teachers employ AI without training in prompting strategies (Lin & Crosthwaite, 2024).

3.3. Procedure

During the setup of the system, the assignment context and the nine-item rubric were devised based on the assignment guidelines provided by the course lecturer (see Appendix A). The items covered the major aspects of the assignment regarding content (items 1 to 4), organisation (items 5 and 6), language (items 7 and 8), and referencing (item 9). Before the system was made available for the course, it had been tested with several writing samples from the previous semester to confirm its capability to generate relevant feedback on all items.

The participants submitted their report drafts to the AI system for feedback. Although some students made multiple submissions, only the final submissions, which students presumably found satisfactory, were analysed. They also received feedback from three of the authors, who were language teachers from the English language centre. After receiving feedback from both AI and teachers, six of the participants, arranged in pairs, took part in three semi-structured online interviews conducted in English by the teachers. These interviews aimed to gain qualitative insights into the participants' perceptions of the AI feedback. The participants were asked open-ended questions about their perceptions of the usefulness of AI feedback, as well as their preferences and strategies for using AI feedback versus teacher feedback. All but one interviewee had their writing assessed by the teacher interviewer. Although having their assessing teacher as the interviewer might have discouraged the interviewees from expressing their honest views, the teacher's experience with the writing task and familiarity with their writing allowed for deeper interaction, which could lead to more meaningful insights.[2] The interviews lasted an average of 28 minutes and were recorded for thematic analysis of the interviewees' experiences with AI feedback. Each teacher interviewer transcribed their respective interviews verbatim with the assistance of the AI-powered transcription function of the online meeting platform. They then checked their own transcripts for accuracy. Table 1 shows the background of the interviewees.

[2] Since 25 % of the report grade was determined by the teacher interviewers, the participants were encouraged during the interviews to give their honest opinions, with reassurance that their grades would not be affected.

Table 1
Background of the interviewees.

Student	Discipline	Year of study	AI tools familiar to the student
12	Biomedical engineering	3	ChatGPT
14	Artificial intelligence and information engineering	2	ChatGPT
20	Civil and environmental engineering	2	ChatGPT, DeepSeek, Grammarly
32	Physiotherapy	2	ChatGPT
36	Occupational therapy	1	ChatGPT, Grammarly
37	Medical laboratory science	1	Poe, ChatGPT, DeepSeek

3.4. Data analysis

The study employed both quantitative and qualitative methods to analyse the AI feedback. The quantitative analysis aimed to evaluate the quality of the AI feedback in terms of accuracy, tone, and provision of examples. The qualitative analysis was conducted to explore the participants' perceptions of the AI feedback.

3.4.1. Quantitative analysis

The AI system generated a total of 333 feedback points from the 37 research report drafts, and they were analysed statistically. Descriptive statistics were used to analyse the distribution of feedback points across two categories: accuracy and tone. Accuracy was included because AI can be susceptible to hallucinations and generate inaccurate feedback (Pack et al., 2024; Steiss et al., 2024). Tone was included because LLMs are often trained to be supportive (Bai et al., 2022; Ouyang et al., 2022; Touvron et al., 2023), which may cause their feedback to favour praise over critical evaluation and limit its usefulness for revision. Since the AI system was instructed to include examples from the report drafts in its feedback, the provision of examples was also examined under the category of accuracy. To identify patterns in the accuracy and tone of AI feedback, the frequency and percentage of feedback points in each category were calculated. The rationale for the classification and the relevant examples are given below.

3.4.2. Classification of AI feedback

3.4.2.1. Accuracy of feedback.

- **Accurate feedback with examples:** Correctly identifies a strength or issue in the draft and supports the judgement with a specific reference to the text (e.g., a quote, paraphrase, or section reference). The cited material accurately reflects the point made.

E.g., *Comprehensive and quality evidence is provided for both the historical background in Section 1 and the critical evaluation in Section 2. The use of sources like Antiquities Advisory Board and Tang (2011) supports the main points effectively, although a broader variety of sources could further strengthen the argument.*

- **Accurate feedback without examples:** Correctly identifies a strength or issue in the draft but does not cite a specific reference to the text.

E.g., *The language used in the report is largely accurate, with few grammatical or syntactical errors. This ensures clear communication of ideas. A closer review for minor errors and a greater variety in sentence structure could further improve readability and engagement.* (Item 8 language, Student 4).

- **Inaccurate feedback with examples:** Incorrectly identifies a strength or issue in the draft and cites a specific reference to the text.

The cited material is irrelevant or misinterpreted and therefore does not substantiate the point.

E.g., *The introduction effectively presents the topic and provides a clear rationale for selecting the topic, as seen in the statement "This essay aims to promote youth awareness of local public housing policy's history and development from the 1950s to the 1990s." However, it lacks a detailed outline of the report's structure, impacting the clarity of what the reader should expect.* (Item 1 introduction, Student 35).

Reason for inaccuracy: The cited statement does not include a clear rationale.

- **Inaccurate feedback without examples:** Incorrectly identifies a strength or issue in the draft but does not cite a specific reference to the text.

E.g., *The report maintains a formal and objective tone throughout, appropriate for an academic setting. The vocabulary is precise, and sentence structures are sophisticated, facilitating an in-depth discussion. However, hedging claims could be used more judiciously to acknowledge potential counterarguments or limitations in the evidence.* (Item 7 tone and style, Student 15).

Reason for inaccuracy: One paragraph is extremely informal.

3.4.2.2. Tone of feedback.

- **Positive feedback**: Highlights strengths or areas of excellence in the writing.

E.g., *The introduction successfully presents Tai Po with essential background, clearly states the research aim, and outlines the structure. It effectively rationalizes the choice of Tai Po as a subject for historical research. For example, "To better understand the history of the New Territories after it was leased to Britain, Tai Po would be a good starting point due to its importance at that time as mentioned above." This shows a compelling rationale for selecting the topic.* (Item 1 content, Student 2).

- **Combination of positive and negative feedback:** Balances recognition of strengths with indication of weaknesses or suggestions for improvement.

E.g., *The first section offers a comprehensive historical perspective, detailing the development of major industries like shipbuilding and textiles. It effectively uses well-cited sources to support the narrative. However, the analysis lacks depth in exploring divergent interpretations or the implications of historical events on Hong Kong's socio-economic landscape.* (Item 2 content, Student 1).

- **Negative feedback:** Focuses solely on identifying weaknesses or areas requiring improvement, without acknowledging strengths.

E.g., *The assignment does not present or evaluate any historical site related to Kwun Tong's industrial history, thus failing to meet the requirements of Section 2 as outlined in the assignment context.* (Item 3 content, Student 31).

3.4.3. Rating procedure and inter-rater reliability

To ensure the reliability of feedback classification, two of the authors with expertise in academic writing reviewed the AI-generated feedback. Both raters were trained using a detailed coding manual that provided definitions and examples for each category. The raters independently classified all 333 AI-generated feedback points for the 37 report drafts, cross-checking each comment against its corresponding draft to judge accuracy and example provision.

Inter-rater reliability was assessed using intraclass correlation coefficients (ICCs) in Statistical Package for the Social Sciences (SPSS Statistics 29.0.2.0). For the category of accuracy of AI feedback, the inter-rater reliability score was 0.899, indicating a high degree of agreement between the raters. For the tone of AI feedback, the inter-rater reliability score was 0.944, demonstrating near-perfect agreement. Discrepancies in classification were resolved through discussion between the raters to ensure consistency in the final dataset.

3.4.4. Qualitative analysis

The three teacher interviewers adopted an inductive approach to thematic analysis by following Braun and Clarke's (2006) framework. Each teacher coded only the interviews they had conducted, thereby taking advantage of the interviewer's familiarity with the interviewees and the context. This direct involvement as interviewer-coder was intended to capture nuanced insights that might otherwise be overlooked. To mitigate any potential biases or misinterpretations arising from each teacher's focus on their own data, they met regularly to discuss and refine their individual codes. Through these discussions, they merged or adapted codes across the three datasets, ensuring that any overlaps were addressed and any points unique to individual transcripts were recognised. Such collaboration led to a unified set of themes, as each teacher's insider knowledge of their own interviews was balanced by iterative reviews of the developing themes by the team.

One interviewee's responses serve as an illustration of how misinterpretations were managed. The interviewee initially appeared to consider AI feedback specific, stating, "*I found that the AI feedback is very specific for me*." (Interview 3, Student 36). This observation was first coded as "AI feedback specific". However, the same interviewee subsequently referred to AI feedback as vague, saying, *"I think AI may give some vague feedback"*, which seemed to contradict her earlier remark and also aligned with the code "AI feedback too vague" used in the other transcripts. A closer examination by the team revealed that the interviewee's notion of "specific" referred to AI feedback being based on the rubric, rather than offering detailed suggestions for improvement. In light of this, the team created the new code "AI feedback targeted" to capture this intent, and the relevant segments were recoded accordingly.

4. Results and discussion

4.1. Quantitative analysis

4.1.1. AI feedback accuracy

Table 2 shows the classification of the 333 AI-generated feedback points according to their accuracy. A large majority of the feedback points correctly identified a strength or an issue. However, over half of them failed to provide specific examples or references to the text to support the observation. As shown in Table 3, when examples were given, the vast majority (81.2 %) focused on items 1 to 4. These items were related to the content of the writing, whereas very few examples were given for the other items and criteria.

Inaccurate feedback accounted for 13.8 % of all feedback points. Of the 46 inaccuracies, nearly two-thirds (63 %) concerned referencing (see Table 4). Typically, the system would assert that the citations were consistently formatted even when they were not or would incorrectly state that a reference list was included. Similarly, inaccuracies about content, organisation, or language may have resulted from AI hallucinations, where the feedback claimed that certain requirements were

Table 2
Distribution of feedback accuracy.

Feedback category	Percentage (N = 333)
Accurate with examples	38.1 % (127)
Accurate without examples	48.1 % (160)
Inaccurate with examples	2.7 % (9)
Inaccurate without examples	11.1 % (37)

Table 3
Distribution of accurate feedback with examples.

Rubric item	Criteria	Percentage (N = 127)
Items 1-4	Content	81.2 % (103)
Items 5-6	Organisation	8.6 % (11)
Items 7-8	Language	9.4 % (12)
Item 9	Referencing	0.8 % (1)

Table 4
Distribution of inaccurate feedback.

Rubric item	Criteria	Percentage (N = 46)
Items 1-4	Content	21.7 % (10)
Items 5-6	Organisation	10.9 % (5)
Items 7-8	Language	4.4 % (2)
Item 9	Referencing	63.0 % (29)

met, such as including an outline, linking paragraphs clearly, or using a personal tone in recounting the site visit, even though these features were clearly absent.

For example, one student included only the aim but not the outline in the introduction yet received an incorrect comment stating that both were given: “The introduction … clearly states the aim of understanding the impact of past decisions on current transport systems and outlines the report’s structure, providing a comprehensive overview as required” (Item 1 content, Student 29). Although the system had an adequately large context window to process the draft and the detailed prompt, hallucinations still occurred. These inaccuracies may be attributable to the next-token prediction training that LLMs have undergone, which prioritises generating plausible outputs based on patterns in the training data rather than ensuring factual accuracy (Banerjee et al., 2025). Consequently, smoothly phrased but occasionally ungrounded feedback can be generated. In addition, since general-purpose LLMs are trained on non-specialised datasets, they may fail to capture the more nuanced aspects of the writing, yielding feedback that is more a reflection of their training than a concrete response to the issues of the writing (Venter et al., 2025). Although researchers are hopeful that more advanced LLMs will improve feedback quality (e.g., Pack et al., 2024; Steiss et al., 2024), hallucinations remain stubbornly challenging to eradicate due in part to LLMs’ current training methods (OpenAI, 2025). As such, it is essential to remind students to engage with AI feedback critically.

Overall, the language of the feedback bore a close resemblance to that of the rubric descriptors, and inaccurate feedback was no exception as it often recycled phrases from the descriptors. For instance, Table 5 presents two cases of inaccurate feedback, where the wording is strikingly similar to the relevant descriptors in the rubric.

This mimicry of the rubric may have been caused by unintended priming from the prompt. Although the system adopted zero-shot prompting (i.e., no examples of student drafts or feedback were provided as demonstrations), the detailed rubric appears to have conditioned the model to treat its phrasing as a high-probability template to guide token prediction. This could be considered conceptually analogous to biases observed in generation tasks with few-shot prompting, where LLMs frequently repeat answers seen in the prompt’s demonstration examples (Zhao et al., 2021). Likewise, including the detailed rubric in the prompt may have biased the model towards paraphrasing or reusing its descriptors rather than providing original critiques. This undesirable effect indicates the need to explore alternative prompting strategies to obtain more specific and accurate feedback (see Section 6).

4.1.2. *AI feedback tone*

As Table 6 shows, the tone of the AI feedback revealed the AI system’s marked inclination towards an apparently balanced approach. An overwhelming majority (90.4 %) of the feedback combined a positive point with a negative one. Very few feedback points were purely positive (9 %) or purely negative (0.6 %). Disregarding inaccurate feedback and focusing only on accurate feedback with examples, 19 (15 %) were positive feedback points, 106 (83.5 %) were a combination of positive and negative feedback points, and 2 (1.6 %) were negative feedback points. A closer scrutiny of the 106 combinations of positive and negative feedback points revealed that the examples were given differently: 65 % of them focused on the positive, 10 % on both, and 25 % on the negative. Taking this into consideration, the tone of feedback with examples revealed a different pattern as seen under “Example focus in feedback” in the table.

This means that when the AI system provided feedback with examples, they were predominantly (69.3 %) focused on the strengths of the writing. Only about one fifth (22.0 %) of the time were examples given to explain weaknesses, and rarely (8.7 %) did the system offer examples of both strengths and weaknesses within the same feedback point. In addition, when the system commented on the weaknesses, it often framed them as suggestions rather than direct critiques (see Appendix B). This bias towards highlighting strengths and softening critiques, despite the system being prompted to exemplify both strengths and weaknesses, might stem from LLMs being trained to produce agreeable responses. During reinforcement learning from human feedback (RLHF), where human trainers evaluate the quality of the models’ responses and provide feedback, LLMs are guided towards respectful and non-confrontational outputs (Bai et al., 2022; Ouyang et al., 2022; Touvron et al., 2023). Because trainers tend to reward polite, supportive language and penalise harsh criticism, the models learned to prioritise praise and hedged phrasing over direct, specific critique, potentially

Table 5
Examples of inaccurate feedback that mirrors the rubric descriptors.

	Inaccurate feedback	Corresponding rubric descriptor	Reason for inaccuracy
Item 6: Organisation	**Each body paragraph** contains well-integrated **evidence that supports the main idea** in the topic sentence and includes insightful analysis that links the evidence back to the main argument, with a **clear progression of ideas.** All are coherently structured with effective **use of connectors**, pronouns, and **summary phrases**. The **overall organisation** of the report also **follows the required structure.** (Item 6 organisation, Student 14)	The report is well-organized, with **each body paragraph** containing **relevant evidence** and analysis **that supports the main idea.** The **progression of ideas is clear,** and the **use of connectors and summary phrases** enhances coherence. The **overall structure follows the required format**, though some sections could be further refined for clarity and flow.	Two body paragraphs do not contain any evidence
Item 9: Referencing	**All sources are clearly acknowledged, and the report effectively summarizes and paraphrases ideas to support the main points. In-text citations and the reference list are formatted accurately according to APA 7th edition style, with at least five sources cited.** For example, ’[1] Yang Zhen, (2013) ….’ accurately follows APA style. However, ensuring consistency in the use of source material and possibly integrating a wider range of sources could further strengthen the referencing. (Item 9 referencing, Student 17)	**The report clearly acknowledges all sources and effectively summarizes and paraphrases relevant ideas to support the main point; all in-text citations and the reference list are formatted accurately according to the APA 7th edition style; at least five sources are cited.**	Both in-text citations and the reference list contain errors

Table 6
Distribution of AI feedback tone and example focus.

	All feedback (N = 333)	Accurate feedback with examples (N = 127)
Feedback tone		
Positive	9.0 % (30)	15.0 % (19)
Combination of positive and negative	90.4 % (301)	83.5 % (106)
Negative	0.6 % (2)	1.6 % (2)
Example focus in feedback		
Examples on strengths[a]	–	69.3 % (88)
Examples on both strengths and weaknesses	–	8.7 % (11)
Examples on weaknesses[a]	–	22.0 % (28)

[a] Includes feedback points that are purely positive or negative and the subset of combination feedback points where the examples pertain to strengths (for 'Examples on strengths') or weaknesses (for 'Examples on weaknesses').

leading them to produce more positive-sounding outputs. Yet, a disproportionate focus on the strengths of the writing may have undermined the practical value of the feedback. As revealed in student interviews, while they were encouraged by feedback that complimented them on their strengths, they also expressed a strong wish for clearer directions for improvement (see Section 4.2.2.1).

4.2. Qualitative analysis

A thematic analysis of the interview data identified five key themes on various dimensions of AI feedback.[3] They were the usefulness and limitations of AI feedback, comparison between AI and teacher feedback, strategies for using AI and teacher feedback together, and suggestions for improving AI feedback. A total of 169 items were coded. Table 7 presents the distribution of the themes and their representative quotes. In the following sections, the five themes will be analysed in light of the quantitative data where applicable, and representative quotes from students' interview data will be given.

4.2.1. Usefulness of AI feedback

4.2.1.1. Encouraging higher achievement and providing insightful ideas. The quantitative analysis of the feedback showed that most of the feedback was accurate, confirming the feasibility of using AI to complement human feedback in writing evaluation (e.g., Banihashem et al., 2024; Escalante et al., 2023). Guided by the assignment context and rubric, the AI system often suggested ways to reach higher performance levels. For example, one feedback point on the report's evidence stated "… integrating more diverse viewpoints could have enriched the analysis" (Item 4 content, Student 23). The assignment context suggested an optional discussion of divergent interpretations of historical events to go beyond simple descriptions. Therefore, this feedback could encourage the student to deepen their analysis critically. The following interview comment demonstrates how students benefited from AI's suggestions for improvement:

> *… AI suggested me to [that I] use more diverse viewpoints to strengthen the analysis … I only provide a descriptive historical information in section one … AI can remind me to use more divergent perspectives, but not just like, maybe Wikipedia …* (Interview 3, Student 36)

In addition, having been trained on vast amounts of data, AI can provide users with distinct insights and perspectives (Benites et al., 2023; Kasneci et al., 2023). The sheer breadth of this training data allows it to propose suggestions that teachers might not have thought of.

[3] The theme "usefulness of teacher feedback" was also identified but since it was outside the scope of the study, it was not discussed in the following sections.

Table 7
Themes identified from the interviews.

Themes	Number of items	Representative quotes
Usefulness of AI feedback	52	*… it told me that … **evidence in my section two could be connected more on how it could engage a modern audience.** … I found that my suggestions were only the ordinary improvements, but I barely mentioned about the modern style and things … So **this is useful.*** (Interview 1, Student 37)
Limitations of AI feedback	53	*And I also think that **AI only points out there are some grammatical mistakes, but … AI didn't specify my grammatical mistakes.*** (Interview 3, Student 36)
Comparison between AI and teacher feedback	36	*I think the **AI feedback comments on very big criteria [general issues]**. But the **teacher's comments focus on the small part and the detail part [specific issues]**.* (Interview 2, Student 14)
Strategies for using AI and teacher feedback together	8	*… **after I watch [read] the AI feedback**, then **I know … the general problem** of my assessment and **the deeper problem I can know by the teacher's feedback**.* (Interview 2, Student 14)
Suggestions for improving AI feedback	20	*I think it would be **better to train the AI system to provide more customised suggestions** for our works. Like, it can point out the exact wording or exact sentence that we need to modify on, like the Turnitin system.* (Interview 3, Student 32)

For instance, in one feedback point about proposing better ways to deliver historical information, AI commented that "… the critique could be deepened by comparing with best practices in museum curation elsewhere" (Item 3 content, Student 21). This was an insightful suggestion, which was not seen among teachers' feedback.[4] Although it was not required by the assignment, such comparison could significantly enrich the analysis. This interviewee was impressed by the notable sophistication that AI displayed in its feedback:

> *So I think it's really kind of smart that it [AI] can accurately find out what kind of drawbacks or some of the points that I can improve into it [on].* (Interview 2, Student 20)

4.2.1.2. Comprehensiveness and efficiency. A key strength of AI feedback identified in the interview data was its comprehensiveness. Since AI offered feedback based on the preloaded assignment context and the rubric, the feedback could address all aspects of the writing as instructed. Conversely, due to time pressures, teachers generally concentrated on the problematic aspects. This aligns with the research findings that AI can outperform human markers in terms of providing feedback that covers a wide range of criteria (Dai et al., 2024; Guo & Wang, 2024) and following the writing genre consistently (Steiss et al., 2024). Students were well aware of this advantage:

> *So, as I say, AI is good. AI provides [a] very, very clear picture on different perspectives [aspects] of my writing, including the language, including different parts, as long as you guys insert the information in it and tell him [it] to train [be trained on] these nine rubrics [rubric items].* (Interview 3, Student 32)

[4] Although teacher feedback was not analysed in the study, it was reviewed whenever AI made an insightful comment or overlooked a major issue. This review aimed to confirm AI's insightfulness and to determine whether the overlooked issue was addressed in the teacher feedback.

Another strength of AI feedback, according to the interview data, lay in its efficiency. Several students highlighted that they could obtain an almost instant response after uploading their drafts for feedback. One student was even pleasantly surprised by its efficiency:

> *… and it is quite fast, actually. I didn't expect it to analyse my entire draft with such speed.* (Interview 1, Student 37)

Efficiency is widely recognised as a main advantage of AI-driven feedback tools (Koraishi, 2023; Li, 2023; Mizumoto & Eguchi, 2023). Perhaps more importantly, beyond convenience, immediate feedback can facilitate students' iterative improvement of writing through multiple rounds of revisions. This allows one of the best feedback practices, providing feedback in multiple stages, to become an active possibility rather than an impracticality due to human constraints (Ferris, 2022). The system's submission records indicated that 13 out of 37 students submitted their drafts multiple times to receive AI feedback. While some of these repeated submissions could have been due to students experimenting with the system, as seen in the short intervals between attempts, others likely reflect students using the feedback to improve their work. This appeared to be the case for those who achieved higher grades in their subsequent submissions. The following comment exemplifies how the immediacy of AI feedback incentivises students to continuously improve their writing:

> *… I can know the feedback immediately and before the deadline, so I can modify my assessment according to the AI feedback to be more [make it] better.* (Interview 2, Student 14)

4.2.1.3. Positive tone. Given that the system was assigned the role of a helpful professor providing feedback on both strengths and weaknesses, the majority of the feedback contained both positive and negative points. This balanced approach aligns with another desirable feedback practice, where feedback should recognise the strengths of the writer, rather than focusing exclusively on the issues in the writing (Goldstein, 2004; Zhang & Hyland, 2022). Students appreciated this approach as they found confidence in the positive points acknowledging the strengths of their writing. The following student comment reflects this view:

> *And I appreciate that it [AI] includes some features that I did [included] in my writing, but it also points out at least … one way to improve my passage. I think that's [it's] important to give some confidence to students that [through indicating] which aspect you did good [well] and what kind of aspects that you can improve.* (Interview 2, Student 20)

4.2.2. Limitations of AI feedback

4.2.2.1. Generic feedback and lack of useful examples. Despite the accuracy and comprehensiveness of AI feedback, in over half of the instances, the AI system did not provide examples or specific references to the text. When examples were given, they were mostly in the form of a positive comment on the sources cited and the sites visited in items 1 to 4 (see Table 3). This was unexpected as the system had been explicitly instructed to provide examples from the text to support its feedback on each item. Apart from the positivity bias that tended to focus examples on strengths, a possible explanation for the system's failure to give examples on all items is the complexity of the prompt, which includes the task instructions, the draft, and the rubric. As found in Pack et al. (2024), AI had more reliable performance with a simpler rubric than one with multiple criteria. Despite the current system's substantial context window, which should have been sufficient to handle the prompt, it still appeared to have reached its processing limit. In addition, since the system was open to concurrent access by students from various courses, it might have processed multiple feedback requests simultaneously. This could lead to system fatigue, resulting in generic feedback that was of limited utility (Awidi, 2024). AI feedback being too vague or general was among the most common themes in the interview data, in which five out of the six interviewees considered some AI feedback too general to be useful. Comments such as these reflect this view:

> *Because AI may only point out where should I [I should] improve, but AI did not give me some specific improvements about how can I [I can] improve this.* (Interview 3, Student 36)

> *I'm not very satisfied [with AI feedback] because, again, it's not very helpful for me. It's too vague and general.* (Interview 1, Student 12)

Given its finite processing capacity and the possibility of system fatigue, the AI system may have resorted to the simple task of scanning for sources and site names and presented them as examples to fulfil the instruction of providing textual evidence. This undermined the usefulness of the feedback and led students to call for more concrete suggestions from AI:

> *I think AI can provide more specific suggestions instead of just pointing out what problems I have in this place [area].* (Interview 3, Student 36)

An extreme instance of the above limitation is seen in the report draft that received the smallest amount of AI feedback. The AI feedback contained only 316 words, appreciably lower than the average feedback length of 453 words. Similar to other generic feedback cases, examples were cited from the draft to support the strengths, whereas none were given to substantiate the weaknesses (see Appendix B).

Besides the possibility of system fatigue, AI's strong tendency (90.4 %) to combine both positive and negative points within the same feedback item also compromised the specificity of the feedback. If the feedback had focused exclusively on either positive or negative points, more text would have been devoted to explaining the strengths or weaknesses. The finding that the majority of the examples (69.3 %) focused on positive points, as illustrated in the extreme case, further reduced the usefulness of the feedback. Although students found positive remarks reassuring, it is evident from the above interview comments that they also value feedback that gives clear directions for improving their writing.

4.2.2.2. Inaccuracies and lack of contextual awareness. Having generic feedback was not the only undesirable outcome of overloading the system with a complex writing task and rubric. Inaccurate feedback, as it turned out, could also be a result of the influence of the rubric. Evidence suggests that AI might produce erroneous comments that can trace their roots to the given rubric (e.g., Pack et al., 2024). As reported earlier (see Table 5), AI tended to recycle phrases from the rubric descriptors in its feedback. One interviewee noticed such inaccurate AI feedback on referencing:

> *It [AI] didn't recognise [the error in] my reference list. It's not alphabetical.* (Interview 2, Student 14)

The inaccurate feedback mirrored the rubric descriptor for referencing, claiming that the report "accurately acknowledges all sources, with in-text citations and the reference list formatted according to APA 7th edition style" (Item 9 referencing, Student 14). Given that the system was set up to provide rubric-based feedback, its feedback would understandably resemble the rubric to some extent. Nevertheless, inaccurate feedback that closely mimics the descriptors, as is the case with other fluent AI-generated texts, could mislead students into trusting its veracity without question (Benites et al., 2023; Gao et al., 2025). This is especially the case for areas such as content and organisation, where students might find it more challenging to detect inaccuracies compared to the technical aspects of referencing. Although the interview data did not cover incorrect AI feedback on these higher-level concerns, it is important to demonstrate these subtle errors. Since content had the second highest incidence of inaccurate AI feedback after referencing, the following two examples illustrate AI's content-related errors.

Choosing public transport as the topic, one student briefly mentioned the Peak Tram's commencement date and key figures, meeting the rubric's surface-level requirements. Then he discussed at length the Tram's capacity and construction materials, concentrating on technical details instead of the required historical significance. This major oversight eluded AI's detection, as the feedback praised the student on the "detailed account of the evolution of public transport in Hong Kong, with specific dates, figures, and events that highlight significant changes over time" (Item 2 content, Student 21). Similarly, when the student made an inappropriate visit to the Hong Kong Railway Museum instead of the Peak Tram Historical Gallery,[5] AI still commended that "the evaluation of the Hong Kong Railway Museum effectively identifies current shortcomings in how information is conveyed" (Item 3 content, Student 21). Neither of the issues was lost on the teacher, who highlighted the gaps and alerted the student to the requirements of the report. This echoes the finding that human markers, with their deeper contextual knowledge, can give more nuanced feedback than AI (Banihashem et al., 2024; Steiss et al., 2024). These major oversights suggest that, even with its vast training data, AI might struggle to consider the broader context when giving feedback (Dai et al., 2024; Farrokhnia et al., 2024), and providing AI with an elaborate rubric could inadvertently steer it towards a parochial focus on surface-level descriptors, causing it to miss the overall picture (Yoon et al., 2023).

4.2.3. Students' comparison of AI and teacher feedback

When comparing AI and teacher feedback, most students expressed a preference for teacher feedback due to its higher quality and relevance. They felt that teachers offered more personalised feedback that better catered to their individual learning and writing needs. Teachers' ability to understand the context of their writing was particularly valued, as it ensured the feedback was more aligned with assignment expectations. The preference for teacher feedback can be seen in the following comments:

> *And it [AI] also didn't include any improvements that I can make … to correct my passage. But for teachers, they did give different comments to help me to know what's the direction that I can improve and what kind of … improvement that I can make.* (Interview 2, Student 20)

> *… if I could only get AI or teacher's feedback, I would rather choose teacher's feedback. Because I think, for [in] my case, I need more customised feedback from teachers.* (Interview 3, Student 32)

4.2.4. Strategies for using AI and teacher feedback together

Although most students preferred teacher feedback for its specificity, they also acknowledged the advantages of AI's speed and accessibility. All of them expressed an interest in a blended approach that combines both forms of support. Several students viewed AI as a preliminary tool to identify basic issues, allowing teachers to focus on deeper, more complex aspects of writing:

> *I think I will first use the AI to generate the feedback first because the teacher's feedback needs some time to generate. And maybe after I watch [read] the AI feedback, then I know … the general problem of my assessment and the deeper problem I can know by the teacher's feedback.* (Interview 2, Student 14)

This blended approach corresponds with the concept of hybrid intelligence (Dellermann et al., 2019; Weber et al., 2025), where artificial and human insights merge to produce results neither could achieve alone. While teachers provide contextual expertise and real-world relevance, AI delivers feedback instantly and at any time. As gatekeepers, teachers also detect inaccuracies or omissions that AI might overlook, thereby ensuring more reliable and substantive feedback for students (Steiss et al., 2024).

4.2.5. Suggestions for improving AI feedback

Most students indicated a need for AI feedback to include specific examples of their writing problems, as they found this type of feedback more actionable and easier to apply to their revisions. One suggestion from the students to achieve this was to integrate an interactive channel into the AI system. This channel would allow them to request clarification or further explanation on the feedback they received:

> *I hope I could … something like ChatGPT, I can point out one specific comment that the AI gave me and ask further.* (Interview 1, Student 12)

Such interactivity would enable students to benefit from a dialogic model of feedback that enhances their understanding and application of the feedback received (Wood, 2021). Considering AI's susceptibility to hallucinations, however, students should be reminded to maintain a critical perspective during their interactions with AI to discern the relevance and accuracy of the feedback (Jiang, 2025; Wang, 2024).

4.3. Summary

In answer to the research question, the findings indicated that the AI feedback system could deliver largely accurate, comprehensive, and efficient feedback that encouraged iterative revisions. The system's customised interface helped generate feedback that covered all the rubric criteria and present it in a consistent format. However, the feedback was often generic with few concrete examples, and some comments were inaccurate. These inaccuracies may have stemmed from AI hallucinations and its tendency to recycle rubric descriptors, together with its inherent difficulty in understanding the broader assignment context. Therefore, despite the system's affordances, it may not be able to meet the nuanced demands of extended academic writing.

Regarding the integration of the detailed rubric into the prompt, it appeared to have mixed effects. The detailed rubric ensured that the feedback covered the full range of writing criteria. Nevertheless, its inclusion might have caused the AI model to recycle the descriptors, and requiring the system to attend to many components could have led to system overload and the resulting formulaic responses. The complexity of the prompt itself also possibly intensified this issue. Furthermore, because of AI's inclination towards positive responses, it tended to overemphasise positive aspects by citing far more examples of strengths than weaknesses in the drafts. These findings indicate that although the carefully designed prompt ensured feedback coverage, balancing the complexity of the prompt with the capacity and tendencies of the AI model remains a challenge.

5. Limitations and directions for future research

This study has several limitations that can be addressed in future research. Constrained by its concurrent use in various courses in the university, the AI system employed zero-shot prompting and was set at zero temperature to minimise randomness in the feedback. Despite the convenience it offers to users untrained in prompting, this design and setting might have prevented AI from utilising its full potential in offering useful feedback. Preliminary evidence suggests that few-shot prompting at higher temperature settings may enable AI to detect language errors more effectively (Xu et al., 2024). By increasing the temperature, AI can sample from a wider distribution of possible tokens, thereby allowing it to generate more varied and nuanced responses. Given the generic feedback provided by the current system, which often closely resembled the preloaded rubric, future research can explore AI-driven feedback systems that allow few-shot prompting at different temperature settings. This could shed light on whether prompting

[5] The two sites have distinct historical significance: The Peak Tram, initially built for wealthy residents such as British government officials, symbolises Hong Kong's colonial past, whereas the Kowloon-Canton Railway was built to connect Hong Kong to mainland China for trade and people movement.

strategies and temperature settings would influence the overall quality of feedback, as well as whether an optimal balance can be achieved between user convenience and feedback usefulness.

The generic feedback without examples suggests that the AI system could have suffered from system fatigue due to high concurrent feedback requests (Awidi, 2024). Ascertaining this, however, requires comparing the feedback from the AI system at varying demand levels. Future research could investigate when the system experiences demand spikes and whether those spikes coincide with higher rates of generic feedback attributable to computational bottlenecks. Confirmation of performance degradation would inform how the system should be optimally employed.

Regarding the interview data, its validity was limited by the small sample size of six students and the fact that the interviews were conducted by the students' own assessing teachers. To obtain more robust support for the quantitative data, future studies should include a larger sample of participants and have interviews conducted by independent researchers who have knowledge of the students' drafts and the two sources of feedback. This should help in obtaining more meaningful and candid responses.

6. Implications for teachers and system designers

The findings have practical implications for both teachers and system designers. Teachers can adopt a hybrid feedback approach that capitalises on AI's affordances while compensating for its limitations (Weber et al., 2025). During early drafting, they can direct students to seek AI feedback for its comprehensive coverage and speedy delivery to support iterative revision. Teachers can then focus their attention on aspects that prove to be challenging for AI: higher-level concerns such as contextual alignment to the assignment and coherence across sections. Given that AI's examples tend to emphasise strengths, teachers should provide concrete guidance when addressing weaknesses. For instance, they could pinpoint specific claims that lack evidence or identify a weak topic sentence and model an improved version. This hybrid approach would enable students to benefit from teachers' actionable feedback in addition to AI's advantages.

The generic and occasional inaccurate AI feedback highlights the need for students' active engagement with the feedback. One classroom practice to turn generic feedback into actionable revision directions is to ask students to give targeted responses. Such responses should include specific revisions they plan to make and quote the exact sentences in their draft that need to be revised. To address feedback that appears to be inaccurate, students can be asked to identify evidence or the lack thereof for any dubious AI comments and give a brief rationale for accepting or rejecting the comments. These activities require students to turn vague suggestions into targeted revisions and critically review fluent but unsupported comments, thereby building their feedback literacy for engaging with feedback independently (Wood, 2021).

Since generic feedback may result from AI's limited capacity to address the entire rubric (Pack et al., 2024), system designers could incorporate flexibility into the system interface by allowing users to select individual rubric criteria for feedback rather than for the entire rubric. For example, students can receive feedback on content for an initial submission, organisation for a subsequent one, and tone and accuracy later. Although this requires multiple submissions to obtain comprehensive feedback on a single draft, the benefit of more detailed feedback justifies the effort. This flexibility also enables teachers to guide students to focus on global issues in early drafts and local issues later in the writing process (Ferris & Hedgcock, 2014). Focusing on content revision first ensures that students' efforts are directed towards developing their ideas, while addressing sentence-level errors later prevents wasted effort on text that could be substantially changed during content revision.

To mitigate AI feedback that recycles rubric language, instead of bundling all inputs into a monolithic prompt, system designers can use chain-of-thought prompting (Wei et al., 2022). This technique instructs the model to break the feedback process into explicit steps within the prompt. For instance, the model can first analyse the writing step by step: identify the objective and rationale, evaluate claims and structure, and review language and referencing. It should then match the results of its analysis with relevant rubric criteria and justify each match. Finally, the model is instructed to offer actionable suggestions based on its analysis and alignment with the criteria and not to copy phrases from the rubric. This guided, sequential process can help the model notice specific strengths and weaknesses and provide tailored suggestions. To reduce the positivity bias, designers can also require the model to prioritise weaknesses and limit praise to a small portion of the feedback. Since chain-of-thought prompting can significantly improve LLM reasoning even in zero-shot settings (Kojima et al., 2022), using it in AI feedback systems may also reduce incorrect feedback due to hallucinations.

7. Conclusion

This study examined a custom-built AI-driven system providing rubric-based feedback on an extended academic report. Despite the AI model's substantial context window and the use of a detailed prompt and rubric, recurring issues persisted: While AI provided largely accurate and comprehensive feedback efficiently, it tended to be generic and lacked specific examples to guide revision. Occasional inaccurate feedback and recycling of rubric language may reflect influences from the model's training and inherent tendencies. More importantly, AI's lack of contextual awareness of the writing task suggests that it may not reliably offer useful feedback on higher-level concerns such as content relevance and coherence. Since AI feedback quality could be constrained by its data processing capacity, users should rigorously assess whether AI-driven feedback systems can manage the complexities of the writing task. Students should also be reminded of AI's fallibility and be encouraged to critically engage with its feedback to avoid developing a false sense of confidence that might arise from its ostensibly relevant and positive comments. This study contributes to the understanding of the affordances and constraints of using a custom-built AI system to provide feedback on extended academic writing. It also offers practical strategies for teachers and system designers to address AI feedback limitations and enhance its specificity and reliability. To ensure the precision and contextual relevance of feedback, we propose a hybrid approach that combines the capabilities of AI with the expertise of human teachers. Future research could explore alternative prompting strategies and various levels of rubric coverage to identify the optimal settings for effective feedback delivery.

CRediT authorship contribution statement

Jim Lo: Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Conceptualization. **Christy Wong:** Writing – review & editing, Writing – original draft, Formal analysis, Data curation. **Agnes Ng:** Writing – original draft, Investigation. **Pinna Wong:** Writing – original draft, Investigation. **Denise Cheung:** Writing – review & editing, Investigation. **Pauli Lai:** Writing – original draft, Software.

Statement on open data and ethics

The datasets used in this study are not publicly available. The authors do not have the authority to share the report drafts, interview transcripts, or recordings; however, other materials are available from the corresponding author upon reasonable request.

Informed consent was obtained from all participants, and their privacy rights were strictly observed.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

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

Acknowledgements

We thank the anonymous reviewers for their valuable comments, which greatly improved the quality of this paper. We are also grateful to Prof. Alice Yin Wa Chan for her insightful feedback on an earlier draft and to the study participants for their contributions.

During the preparation of this work, the authors used GPT-4o to assist with grammar and proofreading. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Appendix A

Assignment context

The research report studies a particular topic in the history of Hong Kong. The research could be related to local histories, e.g., the history of Wan Chai or Tai Po (do NOT choose Central or Sheung Wan), or it could be topical, such as public transport, religion, military or people.

The report consists of two sections:

In Section 1, present a historical perspective on the chosen topic with reference to sources. This section could be largely descriptive but could be analytical if different sources have divergent interpretations of historical events, people or other phenomena.

In Section 2, evaluate how the chosen topic is currently presented at a historical site or sites. Such sites could include a local museum, a monument, or even a plaque. The evaluation should consider how the presentation could be updated, through modern media, to appeal more to a modern audience. This should include an evaluation of the amount and selection of information shared at the site(s).

The APA style (7th edition) should be used. Academic and reliable non-academic sources (e.g., newspapers) are required. Non-English citations must be translated or transliterated into English.

9-item rubric

	Item	Descriptor
1.	Introduction	Introduction: clearly presents the chosen topic and states the theme with essential background information, offers a compelling rationale for selecting the topic, clearly states the aim of the research report and outlines its structure. Please quote examples from the report to support your feedback's content.
2.	Section 1	Section 1: offers a comprehensive historical perspective on the chosen topic, thoroughly describes relevant facts, events, and figures with well-cited sources, and, if applicable, critically analyses and evaluates varying expert interpretations. Please quote examples from the report to support your feedback's content.
3.	Section 2	Section 2: clearly presents an existing historical site or sites relevant to the topic, critically evaluates how information at each site is currently conveyed to visitors, and proposes insightful improvements to enhance the quantity, selection, and delivery of historical content to engage modern audiences. Please quote examples from the report to support your feedback's content.
4.	Evidence	Evidence: offers comprehensive and quality evidence for the historical background in Section 1 and the critical evaluation in Section 2. Please quote examples from the report to support your feedback's content.
5.	Topic Sentence	Topic sentence: clearly and concisely introduces the main idea of each body paragraph, aligning cohesively with the essay's overall thesis. Effectively links with the previous body paragraph as applicable. Please quote examples from the report to support your feedback's content.
6.	Organisation	Organisation: each body paragraph contains well-integrated evidence that supports the main idea in the topic sentence and includes insightful analysis that links the evidence back to the main argument, with a clear progression of ideas. All are coherently structured with effective use of connectors, pronouns, and summary phrases. The overall organisation of the report also follows the required structure. Please quote examples from the report to support your feedback's content.
7.	Tone and Style	Tone and Style: the report employs a formal and objective tone (except in Section 2 where a personal tone can be used for recounting the writer's visit to and impression of a site), using precise vocabulary and sophisticated sentence structures to facilitate an in-depth discussion of the topic, while claims are prudently hedged where necessary. Please quote examples from the report to support your feedback's content.
8.	Language Accuracy	Language Accuracy: the report uses highly accurate language in grammar, sentence structure and vocabulary, ensuring clear communication of ideas without ambiguity. Please quote examples from the report to support your feedback's content.
9.	Referencing	Referencing: the report clearly acknowledges all sources and effectively summarizes and paraphrases relevant ideas to support the main point; all in-text citations and the reference list are formatted accurately according to the APA 7th edition style; at least five sources are cited. Please quote examples from the report to support your feedback's content.

Appendix B

Example of generic AI feedback (Student 35)

1. The introduction effectively presents the topic and provides a clear rationale for selecting the topic, as seen in the statement 'This essay aims to promote youth awareness of local public housing policy's history and development from the 1950s to the 1990s.' However, it lacks a detailed outline of the report's structure, impacting the clarity of what the reader should expect.
2. The essay offers a comprehensive historical perspective with well-cited sources such as Pryor (1972) and Smart (2006), and critically analyzes the evolution of public housing policy. Nonetheless, it could further explore divergent interpretations to enrich its analytical depth.

3. This section clearly presents Mei Ho House as a historical site relevant to the topic and proposes innovative improvements like the use of AI and VR. However, the critical evaluation of current information presentation could be more detailed to better highlight the gaps that the proposed technologies would fill.
4. The report provides comprehensive evidence for both the historical background and the critical evaluation, as seen in the detailed account of the Shek Kip Mei fire and the use of modern technology in museums. Yet, a wider variety of sources could enhance the report's credibility.
5. Most topic sentences clearly introduce the main idea of each paragraph, but some lack a cohesive link to previous paragraphs, slightly disrupting the flow of the essay.
6. The essay is generally well-organized with a logical progression of ideas. However, transitions between some sections and paragraphs could be smoother to enhance coherence.
7. The report maintains a formal tone and uses precise vocabulary throughout. However, some sections could benefit from more sophisticated sentence structures to facilitate deeper discussion of the topic.
8. The language is mostly accurate, but minor grammatical errors and awkward phrasing occasionally hinder clarity.
9. The report cites a variety of sources and mostly adheres to APA 7th edition style. However, inconsistencies in citation formatting and the translation of non-English sources into English are noted.

References

Awidi, I. T. (2024). Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence, 6*, 100226. https://doi.org/10.1016/j.caeai.2024.100226

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv: 2212.08073*. https://doi.org/10.48550/arxiv.2212.08073

Banerjee, S., Agarwal, A., & Singla, S. (2025). LLMs will always hallucinate, and we need to live with this. In K. Arai (Ed.), *Intelligent systems and applications (Lecture Notes in Networks and Systems* (Vol. 1554, pp. 624–648). Springer. https://doi.org/10.1007/978-3-031-99965-9_39.

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education, 21*(1), Article 23. https://doi.org/10.1186/s41239-024-00455-4

Bayerlein, L. (2014). Students' feedback preferences: How do students react to timely and automatically generated assessment feedback? *Assessment & Evaluation in Higher Education, 39*(8), 916–931. https://doi.org/10.1080/02602938.2013.870531

Benites, F., Delorme Benites, A., & Anson, C. M. (2023). Automated text generation and summarization for academic writing. In E. Cotos, C. M. Anson, O. Kruse, A. Shibani, A. Devitt, C. Rapp, & K. Benetos (Eds.), *Digital writing technologies in higher education* (pp. 279–301). Springer International Publishing AG. https://doi.org/10.1007/978-3-031-36033-6_18.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv.2005.14165*. https://doi.org/10.48550/arxiv.2005.14165

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*. https://doi.org/10.48550/arxiv.2303.12712

Chan, C. K. Y. (2024). Students' perceptions of 'AI-giarism': Investigating changes in understandings of academic misconduct. *Education and Information Technologies, 30* (6), 8087–8108. https://doi.org/10.1007/s10639-024-13151-7

Crompton, H., Edmett, A., Ichaporia, N., & Burke, D. (2024). AI and English language teaching: Affordances and challenges. *British Journal of Educational Technology, 55* (6), 2503–2529. https://doi.org/10.1111/bjet.13460

Dai, W., Tsai, Y. S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence, 7*, 100299. https://doi.org/10.1016/j.caeai.2024.100299

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering, 61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*(1), Article 57. https://doi.org/10.1186/s41239-023-00425-2

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research, 83*(1), 70–120. https://doi.org/10.3102/0034654312474350

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education & Teaching International, 61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

Ferris, D. R. (2022). Feedback on L2 student writing: Current trends and future directions. In E. Hinkel (Ed.), *Handbook of practical second language teaching and learning* (pp. 344–356). Routledge. https://doi.org/10.4324/9781003106609-28.

Ferris, D. R., & Hedgcock, J. S. (2014). *Teaching L2 writing: Purpose, process, and practice* (3rd ed.). Routledge.

Fisher, D. P., Brotto, G., Lim, I., & Southam, C. (2025). The impact of timely formative feedback on university student motivation. *Assessment & Evaluation in Higher Education, 50*(4), 622–631. https://doi.org/10.1080/02602938.2025.2449891

Gao, R., Yu, D., Gao, B., Hua, H., Hui, Z., Gao, J., & Yin, C. (2025). Legal regulation of AI-assisted academic writing: Challenges, frameworks, and pathways. *Frontiers in Artificial Intelligence, 8*, 1546064. https://doi.org/10.3389/frai.2025.1546064

Goldstein, L. M. (2004). Questions and answers about teacher written commentary and student revision: Teachers and students working together. *Journal of Second Language Writing, 13*(1), 63–80. https://doi.org/10.1016/j.jslw.2004.04.006

Gozali, I., Wijaya, A. R. T., Lie, A., Cahyono, B. Y., & Suryati, N. (2024). Leveraging the potential of ChatGPT as an automated writing evaluation (AWE) tool: Students' feedback literacy development and AWE tools integration framework. *The JALT CALL Journal, 20*(1), 1–22. https://doi.org/10.29140/jaltcall.v20n1.1200

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining chatgpt's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*(7), 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Hyland, K. (2008). Genre and academic writing in the disciplines. *Language Teaching, 41* (4), 543–562. https://doi.org/10.1017/S0261444808005235

Hyland, K., & Hyland, F. (2019). *Feedback in second language writing: Contexts and issues* (2nd ed.). Cambridge University Press.

Jiang, Y. (2025). Interaction and dialogue: Integration and application of artificial intelligence in blended mode writing feedback. *The Internet and Higher Education, 64*, 100975. https://doi.org/10.1016/j.iheduc.2024.100975

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Khlaif, Z., Odeh, A., & Bsharat, T. R. K. (2024). Generative AI-powered adaptive assessment. In D. Lim, W. A. J. Wan Yahaya, M. Sanmugam, N. M. Mohd Barkhaya, & Z. Khlaif (Eds.), *Power of persuasive educational technologies in enhancing learning* (pp. 157–176). IGI Global. https://doi.org/10.4018/979-8-3693-6397-3.ch007.

Kohnke, L., & Zou, D. (2025). The role of ChatGPT in enhancing English teaching: A paradigm shift in lesson planning and instructional practices. *Educational Technology & Society, 28*(3), 4–20. https://doi.org/10.30191/ETS.202507_28(3).SP02

Kohnke, L., Zou, D., & Su, F. (2025). Exploring the potential of GenAI for personalised English teaching: Learners' experiences and perceptions. *Computers and Education: Artificial Intelligence, 8*, Article 100371. https://doi.org/10.1016/j.caeai.2025.100371

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *arXiv preprint arxiv.2205.11916*. https://doi.org/10.48550/arxiv.2205.11916

Koraishi, O. (2023). Teaching English in the age of AI: Embracing ChatGPT to optimize EFL materials and assessment. *Language Education & Technology (LET Journal), 3*(1), 55–72.

Lee, I. (2008). Understanding teachers' written feedback practices in Hong Kong secondary classrooms. *Journal of Second Language Writing, 17*(2), 69–85. https://doi.org/10.1016/j.jslw.2007.10.001

Lee, D., Kim, H., & Sung, S. H. (2023). Development research on an AI English learning support system to facilitate learner-generated-context-based learning. *Educational Technology Research & Development, 71*(2), 629–666. https://doi.org/10.1007/s11423-022-10172-2

Li, A. W. (2023). Using peerceptiv to support AI-based online writing assessment across the disciplines. *Assessing Writing, 57*, 100746. https://doi.org/10.1016/j.asw.2023.100746

Lin, S., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *System, 127*, 103529. https://doi.org/10.1016/j.system.2024.103529

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics, 12*, 157–173. https://doi.org/10.1162/tacl_a_00638

Ma, Q., Crosthwaite, P., Sun, D., & Zou, D. (2024). Exploring ChatGPT literacy in language education: A global perspective and comprehensive approach. *Computers and Education: Artificial Intelligence, 7*, 100278. https://doi.org/10.1016/j.caeai.2024.100278

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Moorhouse, B. L. (2024). Generative artificial intelligence and ELT. *ELT Journal, 78*(4), 378–392. https://doi.org/10.1093/elt/ccae032

Moorhouse, B. L., Wan, Y., Wu, C., Kohnke, L., Ho, T. Y., & Kwong, T. (2024). Developing language teachers' professional generative AI competence: An intervention study in an initial language teacher education course. *System, 125*, 103399. https://doi.org/10.1016/j.system.2024.103399

OpenAI. (2024, January 25). *New embedding models and API updates*. https://openai.com/index/new-embedding-models-and-api-updates/.

OpenAI. (2025, September 5). *Why language models hallucinate*. https://openai.com/index/why-language-models-hallucinate/.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Proceedings of the 36th international conference on neural information processing systems* (pp. 27730–27744). Neural Information Processing Systems Foundation. https://dl.acm.org/doi/10.5555/3600270.3602281.

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence, 6*, 100234. https://doi.org/10.1016/j.caeai.2024.100234

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing, 57*, 100752. https://doi.org/10.1016/j.asw.2023.100752

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Swales, J. M., & Feak, C. B. (2012). *Academic writing for graduate students: Essential tasks and skills* (3rd ed.). University of Michigan Press.

Tate, T. P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W., & Warschauer, M. (2024). Can AI provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence, 7*, 100255. https://doi.org/10.1016/j.caeai.2024.100255

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*. https://doi.org/10.48550/arxiv.2307.09288

Van Horn, K. R. (2024). ChatGPT in English language learning: Exploring perceptions and promoting autonomy in a university EFL context. *Teaching English as a Second Language Electronic Journal (TESL-EJ), 28*(1). https://doi.org/10.55593/ej.28109a8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 5998–6008). Curran Associates.

Venter, J., Coetzee, S. A., & Schmulian, A. (2025). Exploring the use of artificial intelligence (AI) in the delivery of effective feedback. *Assessment & Evaluation in Higher Education, 50*(4), 516–536. https://doi.org/10.1080/02602938.2024.2415649

Wang, C. (2024). Exploring students' generative AI-assisted writing processes: Perceptions and experiences from native and nonnative English speakers. *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-024-09744-3

Wang, X., Salmani, M., Omidi, P., Ren, X., Rezagholizadeh, M., & Eshaghi, A. (2024). Beyond the limits: A survey of techniques to extend the context length in large language models. In *Proceedings of the thirty-third international joint conference on artificial intelligence (IJCAI-24)* (pp. 8299–8307). https://doi.org/10.24963/ijcai.2024/917

Wang, Y., Zhang, T., Yao, L., & Seedhouse, P. (2025). A scoping review of empirical studies on generative artificial intelligence in language education. *Innovation in Language Learning and Teaching*, 1–28. https://doi.org/10.1080/17501229.2025.2509759

Weber, F., Wambsganss, T., & Söllner, M. (2025). Enhancing legal writing skills: The impact of formative feedback in a hybrid intelligence learning environment. *British Journal of Educational Technology, 56*(2), 650–677. https://doi.org/10.1111/bjet.13529

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv.2201.11903*. https://doi.org/10.48550/arxiv.2201.11903

Wood, J. (2021). A dialogic technology-mediated model of feedback uptake and literacy. *Assessment & Evaluation in Higher Education, 46*(8), 1173–1190. https://doi.org/10.1080/02602938.2020.1852174

Xu, Y., Polio, C., & Pfau, A. (2024). Optimizing AI for assessing L2 writing accuracy: An exploration of temperatures and prompts. In C. A. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 151–174). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.10.

Yoon, S. Y., Miszoglad, E., & Pierce, L. R. (2023). Evaluation of ChatGPT feedback on ELL writers' coherence and cohesion. *arXiv preprint arXiv:2310.06505*. http://arxiv.org/abs/2310.06505.

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

Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arxiv.2102.09690*. https://doi.org/10.48550/arxiv.2102.09690