Exploring GenAI as Evaluative and Formative Assessment Tools in Reading Assessment: A Mixed-methods Analysis of Genre-based Feedback

Chen Ziqi[1], Wei Wei[1*], Chang Sheng[2], Cao Xueyan[1]

[1]Macao Polytechnic University, Macao
[2]Shiqiao Qiaoxing Middle School, Guangzhou, China
[*]Corresponding author's email: weiweitesting@hotmail.com
[*] https://orcid.org/0000-0002-1919-3178

Received: 29/01/2025 Revision: 08/06/2025 Accepted: 09/07/2025 Online: 03/11/2025

ABSTRACT

This study explores the potential of Generative AI (GenAI) chatbots as assessment tools in Computer-Assisted Language Learning (CALL) environments for assessing first-language (L1) reading comprehension, focusing on their effectiveness in providing feedback across three reading genres: classical literature, technical writing, and modern fiction. Using a mixed-methods approach, 360 students' responses to constructed-response items in reading assessments from junior secondary students in China were analyzed, comparing GenAI-generated scores and feedback to those provided by human evaluators. Six expert teachers further assessed the quality of the chatbot's evaluative and revision feedback. Results indicated that GenAI exhibited a significantly stronger alignment with human raters in scoring low-level responses but struggled with high-level samples. Among the genres, interview data suggested that revision feedback for technical writing received the highest ratings for its clarity, rationality, and actionable recommendations. In contrast, feedback for classical literature was often overly complex for junior-level learners and lacked alignment with examination rubrics. For fiction, GenAI struggled with interpretive nuance, thematic complexity, and variability in question types, highlighting its limitations in fostering deep critical literary analysis. This study highlights the genre-specific strengths and limitations of GenAI in supporting reading comprehension.

Keywords: Generative AI, reading assessment, reading comprehension, genre-based feedback

Introduction

The potential of artificial intelligence (AI) to revolutionize teaching and learning environments is immense (Bearman & Ajjawi, 2023; Huang et al., 2024; Chen et al, 2024). Over the past five years AI has been the subject of an increasing body of research on its integration into various aspects of language education (Huang et al., 2023). Previous studies indicated that AI could facilitate language teaching and learning by providing personalised, interactive, and authentic learning contexts (Liang et al., 2023), on the other side, assist teachers in identifying students'

CITATION | Ziqi, C., Wei, W., Sheng, C., & Xueyan, C. (2025). Exploring GenAI as Evaluative and Formative Assessment Tools in Reading Assessment: A Mixed-methods Analysis of Genre-based Feedback. *Computer-Assisted Language Learning Electronic Journal (CALL-EJ)*, 26(4), 378-395.

learning difficulties (Deeva et al., 2020).

A recent systematic review of AI applications in higher education from 2016 to 2022 revealed that language learning was the most common subject area for AI integration, covering writing, reading, and vocabulary acquisition (Crompton & Burke, 2023). The review by Crompton and Burke (2023) identified four primary areas of AI application: (1) assessment, (2) evaluation, (3) predictive analytics, and (4) intelligent tutoring systems. Among these, assessment and evaluation emerged as the most prevalent, with automated assessment tools being the most widely adopted.

LLMs, such as ChatGPT and Deepseek, that have been trained using large amounts of text data can enable Generative AI (GenAI) chatbots to generate responses that are both natural and engaging (UNESCO, 2023; Markowitz, 2023). This capacity enables GenAI chatbots to serve as instructional tools, facilitating student interaction beyond traditional teaching contexts (Tseng & Lin, 2024). Specifically, ChatGPT can reduce teachers' workload by working alongside to help students in language writing classes (Su et al., 2023) and co-design questions (Lee et al., 2023). Moreover, it has the potential to provide constructive feedback on students' text-based assignments, meeting their individual needs (Tseng & Lin, 2024), offering students real-time, detailed support to help them learn, creating opportunities for them to ask follow-up questions and engage in discussion with human educators (Escalante et al., 2023).

Bearing in mind their potential, previous research has highlighted limitations in GenAI chatbots. For example, Wang et al. (2024) indicated that ChatGPT still struggles to fully understand intrinsic logical information in a way that is comparable to humans. As a result, the GenAI chatbot tend to provide generalized and universal feedback to learners in some specific learning areas (Wang et al., 2024). Moreover, to maximize its potential, users need familiar with prompt engineering techniques to effectively utilise and interact with a GenAI chatbot. Without adequate prompt generating skills, the chatbot may not provide desired feedback (Lee et al., 2023; Markowitz, 2023; Nah et al., 2023; Steiss et al., 2024).

Despite the increasing popularity of AI applications and research in higher education, research on LLMs at K12 level, particularly studies on GenAI chatbot's capacity for reading assessment practices, has been scarce. Liang et al.'s (2023) systematic review explored the roles and research foci of AI in language education (AILEd) and included studies published from 1990 to 2020 in the Web of Science database. This review revealed that the research participants were mainly higher education learners, and AI research has been less frequently applied to young learners in K-12 education, and very little has been done from the perspective of teacher instruction. In terms of language skills, most studies focused on essay writing in the tertiary context, whereas listening and pronunciation were the least studied skills. The current study aims to fill this gap by exploring the use of a GenAI chatbot in providing feedback on reading comprehension assessments at the K-12 level.

Literature Review

GenAI evaluative feedback in language assessment

Evaluative assessments generally summarize student performance at the end of a learning period, providing the student with a summative score and limited feedback. Research on the use of LLMs in evaluative assessments has focused on their ability to answer questions (De Winter, 2023; Vázquez-Cano et al., 2023), generate test items (Ghafouri et al., 2024; Lee et al., 2023; Lin & Chen, 2024; Shin & Lee, 2023), and grade answers to compare with human experts'

annotations (Jiang et al., 2023; Lu et al., 2024), particularly in writing and reading assessments.

LLMs' capabilities in summarizing and reviewing academic readings enable these chatbots to demonstrate significant abilities in designing and answering reading assessment tasks (Lee et al., 2023; Lin & Chen, 2024; Shin & Lee, 2023; Vázquez-Cano et al., 2023). For example, a study by Vázquez-Cano and colleagues (2023) showed that ChatGPT-3.5 produced high-scoring responses and with a better writing style than students when answering reading comprehension questions in a large-scale international test. Similarly, Shin and Lee (2023) examined ChatGPT-3.5's ability to produce L2 reading comprehension questions comparable to those created by human experts.

Recent studies have also shown that GenAI chatbots excel in providing revision feedback for writing assessments, such as crafting task prompts, suggesting activities for writing courses and organizing information according to specified rubrics. For example, De Winter's (2023) study revealed that GPT-3.5 and GPT-4 could achieve high composite scores on the Dutch National English Exam, indicating its ability to recognize the genre and respond to writing tasks. Moreover, GenAI chatbots are expected to grade students' academic writing with consistency comparable to human raters, showing a moderate to high degree of alignment with teachers' evaluations (Lu et al., 2024).

GenAI revision feedback in language learning

Revision feedback involves providing language learners with specific suggestions and revision to improve the accuracy, clarity, and overall quality of their written or spoken responses. Recent studies have shown that GenAI chatbots can effectively assist learners with their writing by providing immediate, descriptive, constructive, and personalised feedback (Ghafouri et al., 2024; Tseng & Lin, 2024). For example, researchers have found that GenAI chatbots can enhance or complement instructors' feedback methods through presenting both GenAI- and human-generated feedback (self and peers) (Banihashem et al., 2024; Escalante et al., 2023). A study by Escalante et al. (2023) conducted two longitudinal studies assessing the impact of ChatGPT-4 versus human feedback on student outcomes and preferences. The studies encompassed 43 and 48 students respectively and found no significant difference in learning outcomes and an even split in feedback preferences among the participants. This result indicated that GenAI feedback could effectively supplement traditional teaching methods without compromising educational quality.

While GenAI chatbots are efficient and reliable, they have two notable limitations in language assessment, such as: first, they may struggle to grasp the logical structure of complex arguments, leading to inconsistent feedback quality (Lin & Chen, 2024; Wang et al., 2024); Second, their outputs can be difficult to interpret from learners' perspective and may not always be reliable, raising concerns about their practical use in educational settings (Guo & Wang, 2023). For example, some scholars suggest that GenAI chatbots may be more effective in specific contexts, such as providing formative feedback during the early stages of the writing process (Barrett & Pack, 2023; Steiss et al., 2024). However, Barrett and Pack (2023) examined the perspectives of teachers and students on using GenAI in the writing process. Interestingly, Barrett and Pack (2023) found that teachers have a positive outlook and show openness towards using GenAI tools in the future. Similarly, Ghafouri et al. (2024) believed that integrating GenAI tools into teachers' writing instruction protocols would improve teachers' self-efficacy and assist them in L2 writing instruction.

GenAI in reading comprehension

Reading comprehension refers to the ability of students to understand, interpret, and critically

analyze written texts by understanding literal meanings, synthesizing information, and reflecting thoughtfully on the content (Basaraba et al., 2012). Recent research highlights that AI-based tools can significantly enhance learners' basic reading comprehension, such as extracting information from texts and interpreting implicit meanings through AI's ability to provide immediate feedback, clarify literal meanings, and deliver tailored support (Cheng et al., 2024; Jose, 2024). In the field of academic or scientific papers, GenAI has garnered many supporters (Nguyen-Trung et al., 2024). However, some argue that its accuracy may depend on the academic discipline (Thelwall & Yaghi, 2024), with psychology and neuroscience demonstrating the highest accuracy, while clinical medicine shows the lowest.

While, there is growing evidence that suggests GenAI shows notable deficiencies when applied to higher-order reading comprehension skills such as critical analysis, synthesis of complex ideas, and forming independent judgments (Larson et al., 2024). For example, Altay et al. (2022) demonstrated that chatbots did not outperform traditional methods in shifting medical students' attitudes on complex issues, such as genetically modified organisms, highlighting AI's limited ability to facilitate learners in critical evaluation from diverse perspectives. Sasahara and others (2021) also expressed concerns about AI creating "echo chambers". These are environments where learners encounter only information that supports their existing beliefs, limiting their ability to develop independent opinions. Such challenges are particularly evident in critical reading, where analyzing and questioning diverse perspectives is essential. In addition, while AI can support comprehension at the literal level, Keyes et al. (2021) and Janssen et al. (2022) noted that the over-corrected feedback from AI systems can hinder students' opportunities to develop their contextual skills, leading them to rely too heavily on AI feedback.

Research gap and questions

While much research has focused on writing assessments, shifting the focus to reading comprehension offers an opportunity to gain deeper insights into the strengths and limitations of GenAI technologies in fostering literacy skills among learners. Recent studies, such as Lu et al. (2024) on GenAI in L1 Chinese language assessment and Lin and Chen (2024) on GenAI in reading comprehension question generation, highlight the potential of these technologies in the assessment of L1 writing and reading comprehension. Emerging evidence also indicates that GenAI chatbots have the potential to bridge gaps in traditional teaching methods by providing scalable and timely feedback (Escalante et al., 2023; Ghafouri et al., 2024). However, their capacity to promote critical thinking and deep comprehension remains a subject of debate, particularly when applied to varied reading genres, including classical texts, fictional narratives, and technical writing. Drawing on Halliday and Webster's (2009) Systemic Functional Linguistics, which highlights how genres like classical literature (narrative) and technical writing (expository) differ in linguistic structure and cognitive engagement, we hypothesize that these variations necessitate tailored AI feedback mechanisms to address genre-specific comprehension challenges. To address these challenges, this study aims to fill the gap in the literature by exploring the role of GenAI chatbots in reading assessments across these three genres. The study is guided by the following two research questions:

1. How effectively does Generative AI evaluate students' responses in L1 reading comprehension?
2. How effective is GenAI in generating low-, mid-, and high-level exemplars based on rubrics for L1 reading?
3. What factors influence teachers' perceptions of the effectiveness or limitations of GenAI revision feedback for L1 reading comprehension across classical literature, technical

writing, and modern fiction?

Methods

Context and participants

This study was conducted in a junior secondary school in Canton, an affluent metropolitan area next to Hong Kong. All students in the school own mobile devices, although the school has its own policy to restrict their use in classrooms. Mandarin Chinese is their first language. They also use Cantonese for oral communication and simplified Chinese as their written language. In an examination-oriented educational culture, reading assessments were administered monthly, and language teachers were expected to grade and provide timely feedback on these tests.

Ethical approval was obtained from the junior school headmaster because this study did not involve any students, and the data was primarily secondary data from low-stakes in-house assessments. The researchers were given consent to (1) compare the human scores to GenAI chatbot-generated scores and (2) invite six teachers to evaluate the quality of GenAI chatbot-generated feedback and to participate in follow-up interviews. The six teachers, constituting the full cohort of Chinese language teachers for the studied grade level, were selected for their knowledge of the students' backgrounds, learning requirements, and curriculum assessment criteria. This ensured that the GenAI scoring and feedback assessments were contextually relevant and consistent with the expectations of educational standards.

Material

As part of compulsory education, curriculum and assessment at the secondary level in China focus on equipping students with the skills to communicate in Chinese effectively, appreciate notable cultural accomplishments, and enhance personal growth. The curriculum underscores the importance of being comprehensive and practical and embodying a humanistic spirit, particularly emphasizing practicality (Dong & Xu, 2017). Different genres reflect diverse cognitive and linguistic demands in L1 reading comprehension (Halliday & Webster, 2009). Classical literature requires cultural interpretation and inferential skills, technical writing demands factual accuracy and logical comprehension, and modern fiction involves thematic analysis and emotional engagement. These genres align with the Chinese junior secondary curriculum and high-stakes assessments, ensuring their relevance to learning outcomes such as critical thinking, cultural literacy, and analytical skills, while challenging GenAI's ability to provide accurate and actionable feedback across varied reading demands. For this reason, three genres were selected from a pool of 37 reading passages recommended by the teachers: classic literature from ancient scholars, recently published technical writings from government policy documents, and novellas by a contemporary writer.

The first reading passage was classical literature from the Yuan Dynasty (1279-1368). Based on the chat between the researchers and Ernie, the first reading passage was included in the corpus of the Gen AI Chatbot. The second reading passage was drafted by the participating teachers, drawing on three pieces of recently published government policy documents on promoting local culture and the tourism industry to domestic travellers. Figures, statistics, and promotional flyers were presented in this reading passage to help readers visualize the economic plan. The third reading passage was a short story on homemade rice dumplings, which symbolised the author's childhood life and facilitated an expression of belonging to her family. Table 1 presents the word count of each reading passage. The reading assessment was designed

to measure students' ability to critically evaluate the article and its authentic context. The first author and the participating teachers designed the open-ended questions and marking rubrics. The total mark for each open-ended question is out of 4.

Table 1.

Word length of each reading passage

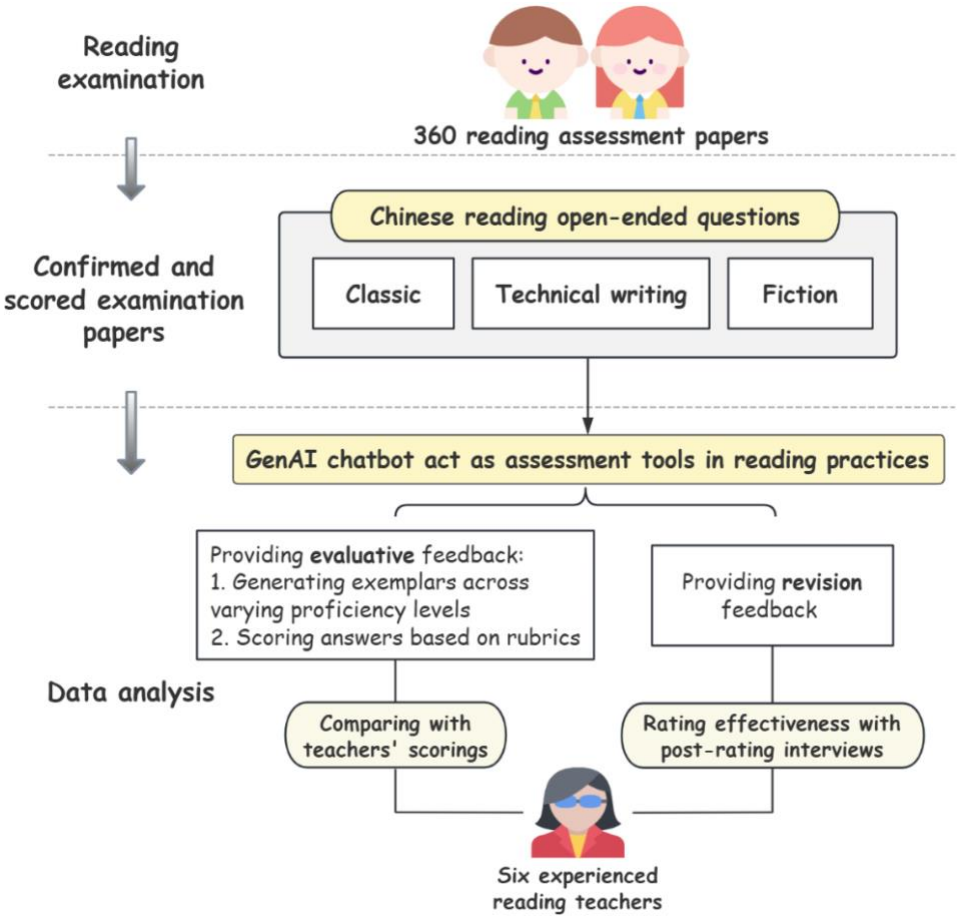
Reading passages	Genre	Included in the AI corpus	Word lengths
1	Classic literatures	Yes	140
2	Technic writings	Partly	662
3	Modern fictions	Not at all	1593

Data collection

The overall data collection procedure of the current study is shown in Figure 3. Specifically, to answer our first and second research questions, we collected a total of confirmed and scored 360 examination papers from a larger pool of student assessments on a centralized server, ensuring data integrity, generalizability, and anonymity. The sample included an equal number of high-, mid-, and low-scored examples (120 responses per level), allowing the GenAI chatbot to grade the papers and generate exemplars at different scoring levels.

Figure 1.

Data collection procedure



The GenAI chatbot employed in this study was Ernie2.0, a ChatGPT equivalent developed by Baidu, which was acknowledged in the recent publication on Generative AI technology in future education by the United Nations Educational, Scientific and Cultural Organization (UNESCO, 2023). It was launched in March 2023 and claimed to excel in various language tasks. Ernie was selected over ChatGPT for several reasons: (1) ChatGPT from OpenAI was not readily accessible to mainland Chinese users due to local school policies prohibiting VPN use on campus, which is required to access ChatGPT; and (2) Ernie was developed using a Chinese corpus and is freely available to both teachers and students. Consequently, its ability to process Chinese reading materials is anticipated to be superior to that of other multilingual Large Language Models (LLMs). In relation to prompt generation, Markowitz (2023) posited the significance of enhancing the quality of feedback by implementing AI prompts that are replete with examples, reflective sequences, and complex cue systems, in contrast to relying on a single prompt or zero-shot prompts. This study adopted a few-shot prompts technique, which involved context setting, providing examples, and executing specific tasks, leading LLMs to generate and produce accurate results.

Table 2.

Prompt phrase for obtaining feedback from GenAI

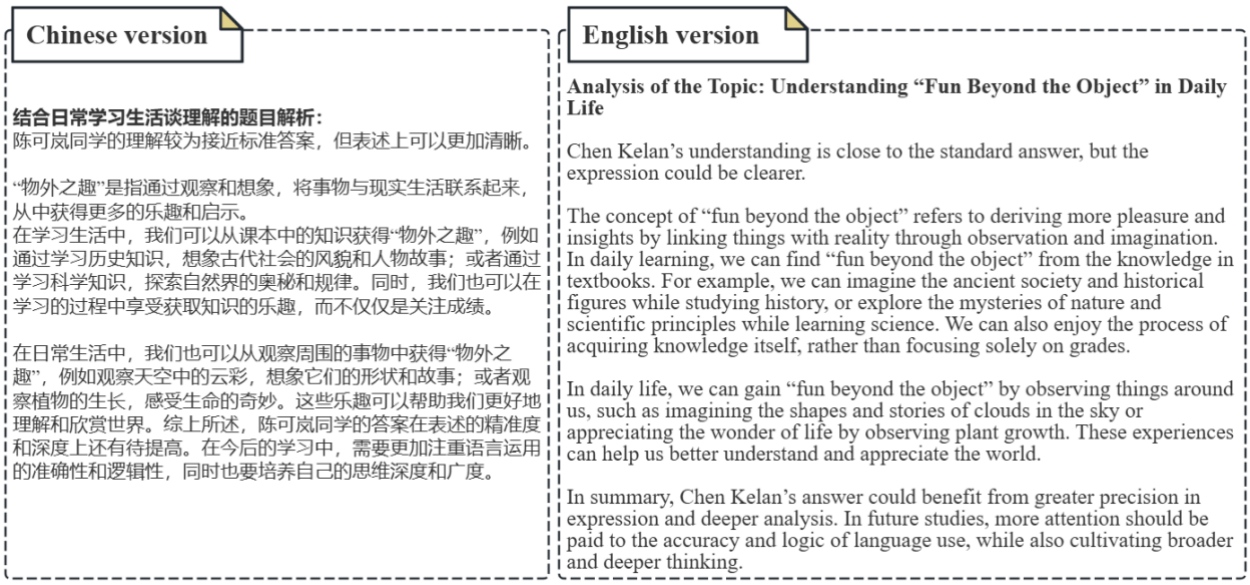
Types of feedback	Prompts
Evaluative feedback: 1. Scoring the answers	English version: Please act as a professional middle school Chinese language teacher. Below are the students' answers (e.g., "I watch birds in the park. It's interesting and not like studying.") Please mark it according to rubrics, assigning scores based on depth of understanding, personal reflection, and text connection.
2. Drafting examples	Please act as a professional middle school Chinese language teacher. Please complete the following tasks step by step: Step 1: Please read the following passage [reading passages]. Step 2: Design the assessment rubrics for three tasks [three open-ended tasks], then draft low-, middle-, and high-scoring answers (e.g., low: minimal response with errors; high: detailed with cultural insight).
Revision feedback: Revised answers based on rubrics	English version: Please act as a professional middle school Chinese language teacher. The following is [student's answer, e.g., "I like playing in the park."]. Please complete the following tasks step by step: Step 1: Noting [model answers provided by the teacher, e.g., "In my daily life, I often find 'the joy beyond material things' when I observe nature during breaks from studying. For example, while walking in the school garden, I imagine the tiny ants carrying leaves as a group of workers building a village, much like the author's imaginative play with mosquitoes and toads. This imaginative perspective makes ordinary moments magical, helping me feel refreshed and find happiness beyond my usual schoolwork."] Step 2: Based on the above information, please revise the student's answer based on [rubrics, e.g., include personal reflection, text connection, and imaginative detail].

The third research question was addressed by using those 360 papers as examples for the GenAI chatbot to provide revision feedback (see Figure 2). The provision of such feedback entails the evaluation of both the content and form of a student's response, with a focus on precision in expression, content depth and analysis, language accuracy and logic, and encouragement of critical thinking. Six expert teachers evaluated the effectiveness of revision feedback by using a questionnaire with a 6-point Likert scale (1 = not useful at all, 6 = very useful). A total of 348 teacher evaluation questionnaires were collected.

Additionally, the teachers provided written responses to follow-up questions (e.g., "*Can you explain the reasoning behind your rating?*"). Interviews with the six teachers were conducted in Chinese, their native language, to ensure natural and accurate responses. The interviews were transcribed and translated into English by a bilingual researcher, with back-translation performed by a second bilingual colleague to ensure fidelity and accuracy of the translated content.

Figure 2.

An example of GenAI-generated revision feedback to classical reading questions



Data analysis

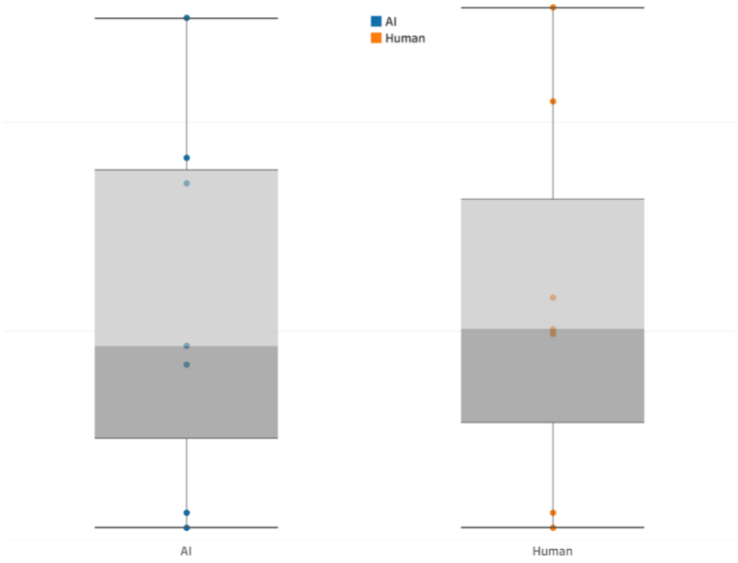
To answer the first research question, AI scores were compared with teachers’ scores through Spearman's rank correlation analysis. A high correlation coefficient may demonstrate AI’s high capability in marking students’ answers. The option coefficient may demonstrate AI’s high capability to marks were invited to mark the three examples provided by the AI. For Spearman's rank correlation, box plots revealed no significant outliers in the distribution of all dependent variables (see Figure 3).

To answer the second research question, we asked the GenAI chatbot to learn the rubrics and generate three answers: high, mid, and low scored. Then, six experienced Chinese teachers were invited to label the three answers blindly without knowing AI’s classifications. Given the discrepancies among the teachers' evaluations, the coordinator of the Chinese language department was invited to discuss with the other teachers to provide a final adjudication. The agreement between AI and teachers may demonstrate the extent to which teachers believe AI

can generate example answers based on marking rubrics at different levels.

Figure 3.

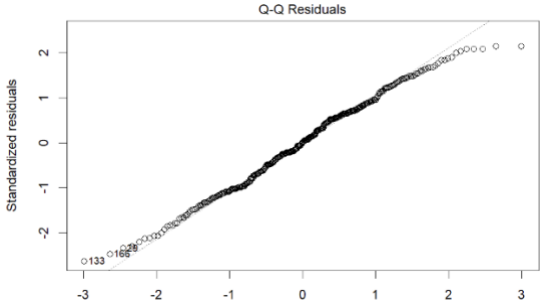
Box plot of the distribution of scores for AI and human teachers



To answer the third research question, teachers were invited to rate revision feedback in a questionnaire on a scale of 6 (1 = not useful at all, 6 = very useful). Controlling students' academic performance, ANCOVA was used to detect possible differences in teachers' average ratings among the revision feedback generated by AI and three types of genres. Several assumptions were tested prior to conducting the ANCOVA. A Chi-Square Q-Q Plot indicated the Multivariate Normality of dependent variables (see Figure 4). If the points lie approximately along the reference line, the data can be considered to be normally distributed. Bartlett's test for the dependent variable indicated that the assumption of homogeneity of variances was met ($p = .583$). Additionally, an analysis of six teachers' responses to a follow-up question in their post-interviews was conducted. To ensure data reliability, the Chinese language department coordinator was invited to assess any discrepancies in the ratings. Thematic coding of interview data was conducted by two researchers to ensure inter-rater reliability, achieving a Cohen's kappa of 0.92. Data saturation was reached after the third interview, as no new themes emerged, confirmed by the fourth interview.

Figure 4.

Q-Q plot for normality



Results

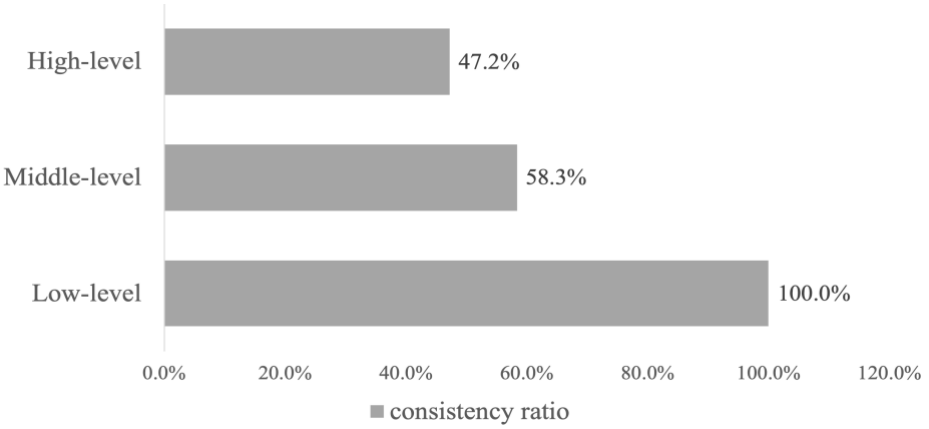
Comparison of GenAI chatbot and teacher assessments

Our first and second research question explores the potential of a GenAI chatbot to provide evaluative feedback that is as accurate and reliable as human teachers in assessment practices. Specifically, we focus on two key capabilities of the AI: first, its ability to respond to open-ended questions in alignment with grading criteria across varying proficiency levels; and second, its capacity to grade student work across three distinct genres of reading passages

To verify the first capability, six expert teachers were asked to evaluate the distinct exemplars generated by the AI, comprising low, middle, and high scores. As illustrated in Figure 5, the results demonstrate that AI and teachers exhibited comparable standards for low-level exemplars, whereas their ratings diverged for high-level exemplars. This discrepancy is likely since the GenAI chatbot lacks the ability to understand complex information and tends to generate responses that are simply copied from the reading material (Lin & Chen, 2024), resulting in significant differences between the high-scoring responses expected by human teachers and those generated by GenAI.

Figure 5.

Consistency ratio of AI and teacher ratings at different scoring levels



Furthermore, we analysed the relationship between scores assigned by the GenAI chatbot and those assigned by human graders using Spearman's rank correlation coefficient. The results are visualized in Figure 2. Table 1 presents descriptive statistics and demonstrates a strong correlation between scores generated by the GenAI chatbot and human raters across all reading passage genres (i.e., *0.75, 0.94, 0.95*). These findings indicate that AI models exhibit a remarkable ability to align with human scoring. Interestingly, despite differences in grading standards, the results suggest a high degree of similarity between GenAI and human raters. The findings suggest that GenAI can be a reliable tool for scoring student responses, enabling teachers to save time on initial grading tasks, especially in large classes. This allows educators to focus on providing in-depth feedback where human judgment is most needed, such as fostering critical thinking.

Table 3.

Descriptive statistics and correlations between grades given by AI and teachers

Type of Reading	Graded by AI		Graded by teacher		*r*
	M	SD	M	SD	
Classic	2.21	0.88	2.01	1.01	.75***
Technical writing	2.33	1.27	2.27	1.34	.94***
Fiction	1.98	1.47	1.80	1.46	.80***

***$p < .001$.

Teachers' evaluation of GenAI chatbot-generated revision

In addition to the capacity of GenAI chatbots to provide evaluative feedback, the third research question posits the question of whether GenAI chatbots effectively provide revision feedback to students in reading assessments. To answer the question, six teachers who had taught the students during the experiment were invited to evaluate the GenAI chatbot-generated feedback. Their classroom teaching experience with these students qualifies them to evaluate the usefulness and quality of the revision feedback generated by GenAI to their students. After collecting their surveys, One-way covariate analysis of variance (ANCOVA) was conducted to test the hypothesis that there would be one or more mean differences between three types of genres (classics, fiction, and technical writings) and GenAI chatbot-generated revision feedback after students' overall reading scores were controlled.

Table 4.

ANCOVA of teachers' evaluation toward revision feedback across genres

Genres	Mean	SD	*df*	*F*	*Sig.*	η^2
Intercept	2.89	1.57	(3,354)	13.12	.000***	0.100
Classic	2.83	1.61				
Technical writing	3.01	1.49				
Fiction	2.86	1.62				

*$p < .05$. **$p < .01$. ***$p < .001$

Controlling for students' grades, the results from the ANCOVA analysis were presented in Table 2, indicating a statistically significant difference between the genres of reading passages and the GenAI Chatbot-generated revision feedback ($F(3,354) = 13.12$, $p < .000$, partial $\eta^2 = .100$). This means that the type of reading genre noticeably affects how well the GenAI's feedback helps students improve their answers. On average, teachers rated the revision feedback for technical writing highest (M = 3.01), indicating it offers clear and actionable suggestions that students can easily apply to enhance their work. Feedback for fiction (M = 2.86) and classical literature (M = 2.83) was slightly less effective but still useful, suggesting it supports students moderately well, likely due to its complexity and misalignment with students' interpretive needs. However, the ANCOVA results did not show significant differences in teachers' evaluations of GenAI feedback quality across all dimensions, indicating that genre-specific variations may be subtle in certain aspects. Therefore, additional qualitative insights are essential for providing an in-depth understanding of teachers' views on their ratings of GenAI feedback across three reading genres: classical literature, technical writing, and fiction.

Within the field of classical literature, teachers identified three significant limitations: 1) impracticality for junior-level students, 2) misalignment with examination rubrics, and 3)

challenges associated with subjective and interpretive elements.

First, teachers highlighted GenAI's limited applicability to junior-level students, as some feedback was deemed overly complex, exceeding the students' comprehension levels and making it less actionable. One teacher stated:

> *In terms of classical literature, its suggestions are relatively strong, but I feel that some aspects still cannot be implemented, as certain parts are relatively difficult for junior-level students.* (Teacher 3)

Second, teachers emphasized that aligning GenAI feedback with standard examination rubrics, rather than solely focusing on student responses, could enhance its relevance and usability. For example, one teacher commented:

> *The problem with AI is that its answer analysis is flawed. It should start from the standard answer key and examination criteria based on scoring points, rather than starting from the student's answer.* (Teacher 2)

Finally, while GenAI excelled in objective tasks, it struggled with more subjective aspects, such as identifying central ideas or providing holistic interpretations of texts. An interview extract illustrates this perspective:

> *The answers in the additional practice questions generated by AI are not complete. For example, it cannot first summarize the central ideas of the text or identify its key elements. From a teacher's perspective, if students know the important points of the text first, they can analyze and answer the questions more effectively.* (Teacher 5)

In the field of technical writing, teachers generally acknowledged the distinct advantages of non-literary-based revision feedback provided by GenAI chatbots. These benefits include: 1) rational and objective feedback, 2) effective analysis with clear suggestions, and 3) concrete recommendations for improvement. For example, as one teacher commented:

> *AI's analysis of students' answers and the responses it generates are clear and understandable for students. Its suggestions are targeted, and if feedback is consistently provided this way, students can gradually form a concept of what is expected. I believe these suggestions are feasible.* (Teacher 1)

For the GenAI revision feedback on fiction, teachers expressed significant reservations about GenAI's performance in fiction genres. Three limitations were identified by teachers, including: 1) lack of interpretive nuance, 2) struggles with complex themes, and 3) provision of vague guidance. Particularly, one teacher commented that GenAI's inability to grasp the interpretive and literary nuances inherent in these texts was a recurring theme. Additionally, teachers noted that GenAI struggled to adapt to the variability of literary question types, often providing vague or misdirected guidance. Its lack of specificity in addressing the unique demands of literary analysis highlighted the need for human intervention to supplement AI-generated feedback. An interview extract illustrates this perspective:

> *Literary fiction question types vary greatly, and with so many variations, you need to guide students in determining what type of question it is, right? Then, you need to think about which aspects to approach the question from and what the essential components of the answer are—these must be clearly pointed out.* (Teacher 6)

Discussion

This study aimed to assess the effectiveness of the GenAI chatbot in providing evaluative and revision feedback on first-language reading assessments across three genres: classical literature, fiction, and technical writing.

Evaluative feedback provided by GenAI

Regarding the evaluative feedback generated by the GenAI chatbot, we found that it exhibited a high degree of similarity to human raters in scoring open-ended questions in reading assessments. This finding aligns with previous research on the effectiveness of GenAI's ability to act as a human instructor to score students' writing tasks (Lu et al., 2024; Jiang et al., 2023).

However, the GenAI chatbot and human teachers demonstrated a high degree of agreement in middle and low-scoring answers, rather than high-scoring answers. This finding is consistent with Steiss et al. (2024), who compared the quality of human-generated and ChatGPT-3.5-generated feedback, indicating that ChatGPT-3.5 lacks the capability to provide high-quality feedback. Consequently, GenAI chatbot-generated high-scoring answers may only replicate the original reading materials rather than generate high-quality responses and insights which require more complex processing (Lin & Chen, 2024; Steiss et al., 2024).

Revision feedback provided by GenAI

In terms of the revision feedback rated by teachers, they rated the effectiveness of GenAI revision feedback for technical writing tasks significantly higher than for classical literature and fiction across the three genres. The results can be explained by the following reasons.

First, teachers' evaluations of GenAI's performance in classical literature indicated significant concerns, mainly focusing on the complexity of the feedback for junior-level students and the misalignment with examination rubrics. This finding aligns with previous research suggesting that AI systems may struggle with tasks requiring a deep understanding of context, subjective analysis, and nuanced interpretation (Steiss et al., 2024; Wang et al., 2024). One of the reasons may be the difficulty in adapting GenAI's feedback to the subjective and interpretive nature of classical texts was a recurring theme in the teachers' comments.

Secondly, the teachers' evaluation on fiction revealed more concerns, particularly in its lack of interpretive depth and struggles with complex themes. This finding echoes with research indicating that AI-driven systems may excel at objective tasks but struggle with the more subjective elements of literary analysis, where context and personal interpretation are critical (Jiang et al., 2023). The reason behind the finding may be GenAI chatbot is unable to cope with various reading tasks in different item types, as they asked test takers to demonstrate a wide range of reading skills. Notably, teachers emphasized the importance of human intervention to supplement this GenAI feedback, specifically when tasks require detailed interpretation and a deep understanding of the text. While GenAI shows promise in providing objective feedback, its ability to engage with higher-order cognitive tasks such as literary analysis needs further development.

Lastly, GenAI's feedback for technical writing was rated more favorably by teachers. This result can be attributed to the rational and objective nature of the GenAI feedback, along with its clear and actionable suggestions, which made it a valuable tool for supporting students' revisions in this genre. In this study, the finding highlights the effectiveness of GenAI in providing feedback on structured tasks, a capability that aligns with AI's strengths in delivering performance-based feedback (Lu et al., 2024). Importantly, teachers also noted that consistent feedback from GenAI could help students gradually internalize what is expected in technical writing,

supporting the development of their reading skills over time.

Conclusion

This study makes three novel contributions to the field of AI in language assessment, particularly in the context of L1 reading comprehension. Firstly, while previous studies have focused on L2 learners or single-genre analyses (Steiss et al., 2024; Jiang et al., 2023), our study uniquely examines the application of GenAI in an L1 Chinese reading context across three distinct genres: classical literature, technical writing, and modern fiction. Secondly, this genre-specific comparison provides a more in-depth understanding of GenAI's strengths and limitations, revealing its effectiveness in objective, structured tasks (e.g., technical writing) while highlighting its challenges with subjective, interpretive tasks (e.g., classical literature and fiction). Thirdly, by employing Ernie 2.0, a model optimized for Chinese language processing, this study extends previous work that primarily utilized models like ChatGPT-3.5, which are less tailored to L1 Chinese contexts. For future improvements, it would be beneficial to focus on enhancing GenAI's ability to interpret complex themes, particularly in literary genres, and to better align its feedback with educational rubrics and standards.

This study has several limitations. First, the small sample size of six teachers, which, while representative of the grade level studied, may not fully capture the diversity of perspectives across broader educational contexts, thereby further constraining the generalizability of the findings. Second, student perspectives are absent, representing a notable limitation, as students are the ultimate users of feedback, and their experiences could provide valuable insights into its effectiveness and applicability. Third, the study did not employ quantitative performance comparisons, such as NLP-based similarity scores, which could have offered a more objective measure of GenAI's alignment with human raters.

Acknowledgments

Financial support for the research was acknowledged by the authors: Macao Polytechnic University Research Grant (RP/FCA-08-2023).

Reference

Agarwal, C., & Chakraborty, P. (2019). A review of tools and techniques for computer aided pronunciation training (CAPT) in English. Education and Information Technologies, 24(6), 3731–3743. https://doi.org/10.1007/s10639-019-09955-7

Altay, S., Schwartz, M., Hacquin, A., Allard, A., Blancke, S., & Mercier, H. (2022). Scaling up interactive argumentation by providing counterarguments with a chatbot. *Nature Human Behaviour, 6*(4), 579–592. https://doi.org/10.1038/s41562-021-01271-w

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). https://doi.org/10.1186/s41239-024-00455-4

Barrett, A., & Pack, A. (2023). Not quite eye to A.I.: student and teacher perspectives on the use of generative artificial intelligence in the writing process. *International Journal of*

Educational Technology in Higher Education, 20(1). https://doi.org/10.1186/s41239-023-00427-0

Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2012). Examining the structure of reading comprehension: do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing*, *26*(3), 349–379. https://doi.org/10.1007/s11145-012-9372-9

Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, 1160-1173. DOI: 10.1111/bjet.13337.

Chen, Z., Zhu, X., Lu, Q., & Wei, W. (2024). L2 students' barriers in engaging with form and content-focused AI-generated feedback in revising their compositions. *Computer Assisted Language Learning*, 1–21. https://doi.org/10.1080/09588221.2024.2422478

Cheng, X., Yin, L., Lin, C., Shi, Z., Zheng, H., Zhu, L., Liu, X., Chen, K., & Dong, R. (2024). Chatbot dialogic reading boosts comprehension for Chinese kindergarteners with higher language skills. *Journal of Experimental Child Psychology*, *240*, 105842. https://doi.org/10.1016/j.jecp.2023.105842

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education, 20*(22), 1-22. https://doi.org/10.1186/s41239-023-00392-8.

De Winter, J. (2023). Can ChatGPT pass high school exams on English language comprehension? *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00372-z

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2020). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education, 162*, 104094. https://doi.org/10.1016/j.compedu.2020.104094

Dong, S., & Xu, J. (2017). Language education in China: the Chinese curriculum. In Encyclopedia *of Chinese language and linguistics*. http://dx.doi.org/10.1163/2210-7363_ecll_COM_00000212

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), 57. https://doi.org/10.1186/s41239-023-00425-2

Ghafouri, M., Hassaskhah, J., & Mahdavi-Zafarghandi, A. (2024). From virtual assistant to writing mentor: Exploring the impact of a ChatGPT-based writing instruction protocol on EFL teachers' self-efficacy and learners' writing skill. *Language Teaching Research*, *0*(0). https://doi.org/10.1177/13621688241239764

Guidance for generative AI in education and research. (2023). In *UNESCO eBooks*. https://doi.org/10.54675/ewzm9535.

Guo, K., & Wang, D. (2023). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*, 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Halliday, M. A. K., & Webster, J. (2009). Continuum companion to systemic functional linguistics. In *Continuum eBooks*. http://www.gbv.de/dms/goettingen/557990661.pdf

Huang, K., Liu, Y., Dong, M., & Lu, C. (2024). Integrating AIGC into product design ideation teaching: An empirical study on self-efficacy and learning outcomes. *Learning and Instruction, 92*, 101929. https://doi.org/10.1016/j.learninstruc.2024.101929

Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Trends, research issues and applications of artificial intelligence in language education. *Educational Technology & Society*, *26*(1), 112-131. https://doi.org/10.30191/ETS.202301_26(1).0009

Janssen, M., Hartog, M., Matheus, R., Ding, A. Y., & Kuk, G. (2020). Will algorithms blind people? The effect of Explainable AI and Decision-Makers' experience on AI-supported Decision-Making in Government. *Social Science Computer Review*, *40*(2), 478–493. https://doi.org/10.1177/0894439320980118

Jiang, Z., Xu, Z., Pan, Z., He, J., & Xie, K. (2023). Exploring the Role of Artificial Intelligence in Facilitating Assessment of Writing Performance in Second Language Learning. *Languages, 8*(4), 247. https://doi.org/10.3390/languages8040247

Jose, J. (2024). The impact of integrating Microsoft Teams – Reading Progress as an Artificial Intelligence (AI) platform for promoting learners' reading aloud skills. *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-13074-3

Keyes, O., Hitzig, Z., & Blell, M. (2021). Truth from the machine: artificial intelligence and the materialization of identity. *Interdisciplinary Science Reviews*, *46*(1–2), 158–175. https://doi.org/10.1080/03080188.2020.1840224

Larson, B. Z., Moser, C., Caza, A., Muehlfeld, K., & Colombo, L. A. (2024). Critical thinking in the age of generative AI. *Academy of Management Learning and Education*, *23*(3), 373–378. https://doi.org/10.5465/amle.2024.0338

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2023). Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-12249-8

Liang, J., Hwang, G, Chen, M., Darmawansah, D. (2023). Roles and research foci of artificial intelligence in language education: An integrated bibliographic analysis and systematic review approach. *Interactive Learning Environments, 31*(7), 4270-4296. https://doi.org/10.1080/10494820.2021.1958348

Lin, Z., & Chen, H. (2024). Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. *System*, 103344. https://doi.org/10.1016/j.system.2024.103344

Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? *Assessment & Evaluation in Higher Education*, 1–18. https://doi.org/10.1080/02602938.2024.2301722

Markowitz, D. M., & Hancock, J. T. (2023). Generative AI Are More Truth-Biased Than Humans: A Replication and Extension of Core Truth-Default Theory Principles.

Journal of Language and Social Psychology, 43(2), 261-267. https://doi.org/10.1177/0261927x231220404

Nah, F. F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research, 25*(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814

Nguyen-Trung, K., Saeri, A. K., & Kaufman, S. (2024). Applying ChatGPT and AI-Powered tools to accelerate evidence reviews. *Human Behavior and Emerging Technologies*, *2024*(1). https://doi.org/10.1155/2024/8815424

OpenAI. (2023). GPT-4 Technical Report. *arXiv Preprints*. https://arxiv.org/abs/2023.08774

Sasahara, K., Chen, W., Peng, H., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2020). Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, *4*(1), 381–402. https://doi.org/10.1007/s42001-020-00084-7

Shin, D., & Lee, J. H. (2023). Can ChatGPT make reading comprehension testing items on par with human experts? *Language Learning & Technology, 27*(3), 27–40. https://hdl.handle.net/10125/73530.

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

Thelwall, M., & Yaghi, A. (2024, September 25). *In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results*. arXiv.org. https://arxiv.org/abs/2409.16695

Tseng, Y., & Lin, Y. (2024). Enhancing English as a Foreign Language (EFL) Learners' Writing with ChatGPT: A University-Level Course Design. *Electronic Journal of e-Learning*, 00. https://doi.org/10.34190/ejel.21.5.3329

Vázquez-Cano, E., Ramírez-Hurtado, J. M., Sáez-López, J. M., & López-Meneses, E. (2023). ChatGPT: The brightest student in the class. *Thinking Skills and Creativity*, *49*, 101380. https://doi.org/10.1016/j.tsc.2023.101380

Wang, L., Chen, X., Wang, C., Xu, L., Shadiev, R., & Li, Y. (2024). ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study. *Thinking Skills and Creativity*, *51*, 101440. https://doi.org/10.1016/j.tsc.2023.101440

Xu, Z., Wijekumar, K., Ramirez, G., Hu, X., & Irey, R. (2019). The effectiveness of intelligent tutoring systems on K-12 students' reading comprehension: A meta-analysis. *British Journal of Educational Technology, 50*(6), 3119–3137. https://doi.org/10.1111/bjet.12758.

Biodata

Chen Ziqi is a PhD candidate in Educational Technology and Innovation at the Faculty of Applied Sciences, Macao Polytechnic University. She holds an MSc in Data Science from the University of Macau. Her research interests focus on AI in education and technology-enhanced practices in educational assessment and evaluation.

Dr. Wei Wei is an Associate Professor at the Faculty of Applied Sciences, Macao Polytechnic University. He has a PhD from the School of Education, University of Leeds, UK. His research interests include computer-assisted teaching and learning and learning-oriented assessment practice.

Chang Sheng is a secondary teacher at Shiqiao Qiaoxing Secondary School, Guangzhou, Guangdong, China. He holds a Master of Education in Educational Leadership from the Education University of Hong Kong, has more than five years of teaching experience, and his research interests include innovative teaching methods and educational technology integration.

Cao Xueyan is a PhD candidate in Educational Technology and Innovation at the Faculty of Applied Sciences, Macao Polytechnic University. She earned her master's degree from China University of Mining and Technology. Her research focuses on AI-driven learning systems and intelligent learning technologies.