《外國語文研究》第四十二期
2025 年 12 月 頁 107-134
DOI:10.30404/FLS.202512_(42).0005

研究人工智慧聊天機器人在香港大學英語必修課中自動進行寫作評估的潛力

李禕 *、楊潔 **、何璿子 ***、蘭舸 ****

香港城市大學

摘要

自從 2022 年 ChatGPT 發佈以來，生成式人工智慧在語言教育的多個方面產生了巨大影響，例如語言教學、習得和評估。以香港某公立大學爲例，本研究旨在探究 ChatGPT 是否可以作爲必修英語課程的寫作評估工具。爲了在這一教學情境中實現生態有效性，本研究構建了一個由人工智慧驅動的聊天機器人，以模擬英語教師在評估學生寫作任務前需要經歷的完整培訓過程，例如閱讀作業要求、熟悉評估標準以及審閱學生寫作樣本。該聊天機器人被用於自動評估 100 篇由母語爲中文的本科生產出的敘述性作文。研究結果顯示，在三個總體寫作質量等級（A、B 和 C）之間存在輕微一致性（Kappa 值 $\kappa = 0.126$），而機器人評分與教師評分之間存在正向中度相關（$r = 0.446$）。該結果揭示了在英語課堂中使用生成式人工智慧自動評估寫作的機遇與挑戰。最後該研究提出了對未來類似研究和語言教學實踐的啓示。

關鍵詞：寫作評估、生成式人工智能、第二語言寫作

* 香港城市大學陳馮曼玲陳淑蓮語言中心講師
** 香港城市大學英文系博士生
*** 香港城市大學英文系研究助理
**** 香港城市大學英文系助理教授

Investigating the Potential of a Customized AI Chatbot to Automate Writing Assessment in a Compulsory English Course at a Hong Kong University

Yi Li*, Jie Yang**, Xuan-Zi He***, Ge Lan****

City University of Hong Kong

Abstract

Since the release of ChatGPT in 2022, Generative AI has brought a large influence on multiple aspects of language education, such as teaching, learning, and assessment. This study aims to explore whether ChatGPT can be used as an assessment tool in a compulsory English course at a Hong Kong university. To achieve high ecological validity in this teaching context, an AI-powered Chatbot was built to replicate the exact training process that English teachers need to undergo before assessing students' writing tasks, such as reading assignment prompts, familiarizing themselves with the assessment rubric, and reviewing standardization samples. This Chatbot was applied to automatically score 100 narrative essays written by undergraduate students, who speak Chinese as their first language. The findings show a slight level of agreement (Kappa value of $\kappa = 0.126$) between three general grade levels (A, B, and C) and a positive, moderate

* Instructor at Chan Feng Men-ling Chan Shuk-Lin Language Centre, City University of Hong Kong

** PhD candidate of the Department of English, City University of Hong Kong

*** Research Assistant of the Department of English, City University of Hong Kong

**** Assistant Professor of the Department of English, City University of Hong Kong

correlation (r = 0.446) between Chatbot scores and teacher scores. These findings reveal both the opportunities and challenges of using GenAI to automate writing assessment in English classrooms. The study concludes with implications for future research and language teaching practices effectiveness of GenAI in this context.

Keywords: writing assessment, Generative AI, second language writing

1. Introduction

Generative Artificial Intelligence (GenAI) has gained significant attention in education since the release of ChatGPT in 2022. A growing body of research has emerged to explore the potential in applying GenAI in language teaching and learning. In large public universities in Hong Kong (HK), one notable challenge faced by language teachers is marking student writing and provide immediate feedback. This is due to the large student population in compulsory English courses. The advent of GenAI brings possibilities to mitigate such challenge: teachers could outsource some of the instructional duties to GenAI and redistribute more time and effort to other aspects of teaching (e.g., teaching innovation, materials development). One of the main areas that has garnered considerable interest is the application of GenAI for automated essay scoring (AES), given its advanced reasoning capabilities (e.g., Bui & Barrot, 2024; Geçkin et al., 2023; Mizumoto & Eguchi, 2023). As a new research focus, however, mixed findings have been found on the effectiveness of using GenAI tools to perform scoring comparable to human raters, necessitating more follow-up investigations for this line of research. One plausible explanation for the inconsistency in results is that GenAI may inherit biases from its training data when it comes to assessing student writing from different contexts. To further advance the research of using GenAI for automated writing assessment, we conducted this study with academic writing produced by Chinese students from a mandatory English course in a government-funded university in HK. This study aims to explore the relationship between a) the total score of an in-class timed writing task given by English instructors and b) the total score generated by a customized GenAI Chatbot. The findings would contribute to our understanding of the possibilities and limitations of using GenAI tools for writing assessment in this or similar English courses.

2. Literature review

2.1. The Role of GenAI in Writing Research

ChatGPT was released in November 2022, marking a milestone in the advancement of GenAI. The generative capability of GenAI is driven by its large language models (LLMs), which are based on sophisticated machine learning algorithms to simulate the process of classification, prediction, and decision-making, similar to human brains. As Mizumoto and Eguchi (2023) explained, GenAI tools, e.g., ChatGPT, are trained with a large amount of data enabling them to achieve an array of language tasks, such as text generation, text classification, and language translation. AI is not a

new term in academic communities, since it has been discussed for several decades and has received substantial research attention.

However, only a few scholars in applied linguistics integrated AI into their research previously, perhaps because this integration requires specialized skills in computer science and a strong statistical knowledge foundation. Nevertheless, GenAI tools open an expansive venue for people without such advanced skills to access and apply advanced built-in algorithms, leading to an unprecedented influence on nearly all possible disciplines (e.g., science, social sciences, humanities). These tools include but are not limited to ChatGPT by OpenAI, Claude by Anthropic, Copilot by Microsoft, and Gemini by Google, to name a few.

Pertaining to writing studies, four main themes have been identified: 1) the assistance of AI to aid the writing process (e.g., Kim et al., 2004; Guo et al., 2004; Su et al., 2003); 2) the potential of amalgamating teacher feedback with AI-generated feedback (e.g., Zhang et al., 2024; Guo & Wang, 2023; Su et al., 2023); 3) using AI to enhance writing motivation (e.g., Kim et., 2024; Lu et al., 2024; Teng et al., 2024); 4) leveraging AI for essay scoring (e.g., Geçkin et al., 2023; Mizumoto & Eguchi, 2023; Shin & Lee, 2024). With the arrival of ChatGPT, writing assessment tools have expanded their functions from mere proofreaders or grammar checkers to becoming well-integrated into the entire writing process. Numerous scholars, such as Kim et al. (2024), Guo et al. (2024), and Su et al. (2023), have acknowledged GenAI's ability to scaffold students in their writing tasks, ranging from outlining, revising, editing, and proofreading tasks, within writing classrooms. In a study exploring students' perspectives on GenAI-assisted academic writing, students valued the positive role of GenAI in two particular aspects, which are generating initial ideas for elaboration and structuring content in a logical progression (Kim et al., 2024). In addition, GenAI tools have proven instrumental in providing feedback on student writing. Given the labor-intensive nature of teacher feedback, GenAI has been widely recognized as a supplementary tool in classroom settings, offering suggestions for revisions on various writing dimensions, including language, content, and organization (Zhang et al., 2024; Guo &Wang, 2023; Su et al., 2023), while teachers tend to be oriented towards content and language. However, Su et al. (2023) noted that compared to teacher feedback, feedback given by GenAI tools such as ChatGPT is sometimes general, requiring further clarification. This limitation is echoed by Lu et al. (2024). Still, they argue that ChatGPT's feedback may play a pivotal role in encouraging students' critical thinking, ultimately leading to their self- initiated new writing revisions. Further,

some studies have also revealed increased student engagement in the writing process. For example, Kim et al. (2024) summarized how GenAI enhances the affective domain. First, students have reported experiencing joy and excitement when receiving immediate answers from ChatGPT, transforming the solitary writing task into a collaborative and interactive experience. The constant presence of AI offers a sense of relief and companionship for students. This psychological support tremendously benefits second language learners who often struggle to complete writing assignments because of language barriers. Additionally, ChatGPT facilitates inquiry-based writing instruction by promoting students' ability to ask effective questions to ensure high-quality output from ChatGPT.

2.2. Recent studies of using GenAI for Automated Writing Assessment

A recent area of study is the use of GenAI in automated writing assessment. Compared to human rating, which can be time-consuming and labor-intensive, GenAI tools can efficiently offer substantial assistance in writing assessments (e.g., automated scoring). Although a consensus on the reliability of GenAI in automated writing assessment has not yet been universally achieved, it is widely acknowledged that GenAI can be potentially employed for this purpose. Having said this, its application needs to be under prudent guidance to harness its potential advantages, such as expedited rating and enhanced objectivity. The majority of the recent studies (e.g., Geçkin et al., 2023; Mizumoto & Eguchi, 2023; Shin & Lee, 2024; Yamashita, 2024) reported a general agreement between GenAI scores and human scores, confirming the reliability and utility of GenAI tools in automated essay scoring. However, a few exceptions (e.g., Bui & Barrot, 2024; Shabara et al., 2024; Yancey et al., 2023) revealed low or varying degrees of agreement of the scores provided by GenAI and experienced human raters. The following is a review of key studies investigating GenAI's reliability as an AES tool influenced by different factors: 1) the rubric type; 2) the assessment genre; 3) student populations; 4) the ChatGPT version and its customization.

Geçkin et al. (2023) employed ChatGPT 3.5 to automatically score analytical essays produced by 43 university students who are advanced English learners in Turkey. They compared the AI scorings using a holistic rubric with their counterparts from human raters, showing a significant but weak correlation via Spearman correlation ($\rho = 0.237$). There was also a slight to a fair level of agreement between ChatGPT 3.5 and the average scores of the five raters (ranging from $\kappa = 0.027$ to $\kappa = 0.26$ for five

different human raters). Tate et al. (2024) extended the research scope by comparing two versions of ChatGPT (3.5 vs 4.0) regarding their scoring accuracy while still using a holistic rubric. A large corpus of academic essays produced by secondary school students from primarily English- and Spanish-speaking countries was used. They found that ChatGPT 4 achieved stronger internal consistency, "with GPT-4 in exact agreement with itself over 80% of the time, compared to GPT 3.5 (approximately 60%) and humans (43%)" (p.6, Tate et al., 2024). Their ANOVA test showed no significant difference between the human scores and the scores from the two versions of ChatGPT, but they concluded that ChatGPT 4 tended to assign higher scores than humans, and the scores assigned by both ChatGPT series were reserved, lacking extraordinarily high or low scores.

Mizumoto and Eguchi (2023) investigated the potential of using ChatGPT 3.5 to assess the high-stakes international language proficiency test TOEFL iBT. They analyzed 1210 argumentative essays written by second language learners with diverse linguistic backgrounds (e.g., Chinese, Hindi, Spanish). The results showed some variation between the scores given by ChatGPT and the benchmark TOEFL scores, typically showing a 1-2 point difference, with most discrepancies being just 1 point on a publicly available 10-point rubric for the IELTS Writing Task two. The exact agreement was low (54.33%), but the adjacent agreement was substantially strong (89.15%). The study concluded that the ChatGPT-generated scores generally reflected the three writing levels of TOEFL (i.e., low, mid, high), indicating its possibility for use in AES.

Then, a more recent study by Bui and Barrot (2024) experimented with an analytic rubric on 200 argumentative essays written by Asian students from different countries and regions (e.g., Korean, Chinese, Japanese, and Indonesian) with varying English proficiency levels (i.e., from low to high). The findings cast some doubt on ChatGPT 3.5's independent capability as an AES tool. The results of the Pearson correlation analysis revealed a weak relationship across five assessment domains (e.g., audience, cohesion, and language conventions) and the overall score (ranging from $r = 0.172$ to $r = 0.393$). The intraclass correlation (ICC) was also calculated to measure the internal consistency of ChatGPT 3.5 at two different time points. The low ICC values implied poor reliability in both the overall score and across the five domains (ranging from $r = 0.295$ to $r = 0.481$). However, it is worth noting that despite the preliminary research results, Bui and Barrot pointed out that as the LLMs are constantly improving with more customized training data, the GenAI tools will play a significant role in future writing instruction and

assessment.

Following the call for calibration of scoring standards, Shabara et al. (2024) examined the consistency and accuracy of ChatGPT 3.5 by presenting it with different samples that are representative of different bands of the 0-100 point analytic rubric (10 bands in total) first before outputting scores. 100 expository writing samples from EFL undergraduate students in Egypt were analyzed. Regarding internal consistency, the results suggested that ChatGPT exhibited a moderate level of intra-rater reliability (rICC = .69, $p < .01$, 95% CI [.54-.79]).

The inter-rater reliability showed moderate agreement between ChatGPT and the teachers on some assessment domains (i.e., Communicative Quality, Use of Academic Vocabulary and Style) but poor agreement on others (i.e., Organization, Content, and Relevance). Overall, there was a weak inter-rater reliability between their final scores (rICC = .47, $p < .01$, 95% CI [.00—.70]). Shabara et al. hypothesized that the overall low score reliability might be attributed to ChatGPT's more varied scoring distribution compared to teachers.

To the best of our knowledge, so far, only Shin and Lee (2024) have utilized the customized Chatbot, i.e., *My GPTS*, based on ChatGPT 4 for AES. Prior to scoring, the instructions for the Chatbot, the scoring rubric, and sample essays were uploaded. The assessment, using a 1-5 point analytic scale, included 50 argumentative essays of 80-120 words in length written by Korean secondary school EFL students. The results of the intraclass correlation coefficient showed a strong resemblance in scores given by both the ChatGPT-4 based Chatbot and human raters (r=.91 for Organisation, r=.93 for Language Use, r= .95 for Task Completion and Content). Yet, they noted that the Chatbot generally awarded higher scores than human raters in most domains, except for Language Use.

2.3. Gaps and Research Questions

After the review of existing studies, we identified three important research gaps. The first gap is that, although several studies have been conducted to explore using GenAI tools for automated essay scoring and/or assessment, this line of research is still underexplored because GenAI is a fairly new focus in recent studies. Scholars have noted that GenAI is undergoing rapid changes and improvement (e.g., Bui & Barrot, 2024). Its computing power is expected to be continuously elevated with more training data and users. The second gap is that most studies fall into two main types of academic writing tasks, including argumentative and expository essays; however, research on other common academic genres (e.g., narrative essays)

remains inadequate, which limits our understanding of how GenAI can facilitate writing assessment in tertiary education. The third gap is that, as far as we are aware, few studies have replicated the exact procedures that teachers need to follow before scoring students' writing tasks. In other words, no customized Chatbots have been implemented for AES purposes except for Shin and Lee (2024); however, their study was conducted in a secondary school setting with a relatively small sample size. Thus, further investigation is needed to help us achieve a more comprehensive understanding of how effective a customized Chatbot can be in assessing underexplored but common writing tasks in a university setting.

To bridge these gaps, we collected a writing task in an important genre (i.e., narrative essays) from a mandatory English course provided by a public university in HK. Additionally, based on ChatGPT-4o (under the Poe premium subscription), a Chatbot was developed with prior training on the instructional materials of the course, the assessment itself, and standardization samples of each grade level. The Chatbot was instructed to automatically assess students' narrative essays. Using the rubric that has been consistently implemented in the actual course for years, it generated total scores for the narrative essays along with justification notes. This study aims to answer two research questions:

1. What is the level of agreement between the Chatbot scores and the teacher scores on assessing the narrative essays of university students?
2. How are the Chatbot scores correlated with the teacher scores on assessing the narrative essays of university students?

3. Method

3.1. Setting and participants

This study was conducted in a government-funded university in HK, where English is used as the medium for instruction. The data in our study was collected from the English for Humanities and Social Sciences course, and it is mandatory for undergraduate students who study in the College of Liberal Arts and Social Sciences. Students taking this course are expected to achieve Grade 4 (i.e., equivalent to 6.5-7.0 in IELTS) on the English subject of the Hong Kong Diploma of Secondary Education (HKDSE) or have successfully completed prerequisite English for Academic Purposes (EAP) courses. The course is designed to cultivate students' academic literacy skills through a number of activities, including critical reading, awareness of the key characteristics of key social science and humanities genres, and writing subskills, such as supporting an argument with effective evidence.

The assessment task was an in-class timed narrative essay designed to help students develop an in-depth knowledge of how humans perceive and interpret the world and their experience in relation to time and place. The ability to decode and articulate human experience into temporally meaningful episodes is essential for students in disciplines such as psychology, anthropology, linguistics, sociology, and philosophy (McAdams, 2008; Özyıldırım, 2009). Students are expected to reference Labov (1972)'s narrative model when composing their essays. The narrative essay should follow a structured format, incorporating different components, such as orienting readers to the story using classic wh-questions, introducing a complicating action that creates conflict, describing how the conflict is resolved, and concluding with a statement about future actions. A prompt is released in class on the exam day, with topics pertinent to their university lives so that students can draw on their personal experiences. For example, in the imagined scenario of applying for an overseas summer volunteer program, students may be asked to write a personal narrative about a time they made a positive impact in their community and explain the benefits of this experience in terms of community engagement. The exam is usually conducted during Week 9 of the 13- Week academic semester. Students are required to use the university's lockdown browser to complete their writing, and other external tools (e.g., Grammarly, ChatGPT, or other GenAIs and dictionaries) are strictly prohibited during the exam period. Students are given 2 hours to produce an essay of approximately 600-800 words. No late submissions are permitted. The essay weighs 40% of the total course grade.

The data was collected from 2022 to 2024. Initially, 117 participants (i.e., university students) consented to participate in the study. 17 participants, however, were excluded at a later point for several reasons, including the absence of their individual domain scores on this assessment, cases of plagiarism, and failure of the final course grade. Therefore, a total of 100 narrative essays were included in the final data analysis. The first language the students used was Chinese, including Cantonese (72), Mandarin (12), and both (16). As the course mainly targets freshmen students with an aim to support their English language ability for their majors, the majority of the students (87) were first-year undergraduates. A smaller group (13) consisted of second-year or above students, typically due to circumstances such as holding an associate degree of a 2-year program from other universities or completing the pre-requisite EAP courses. 38 of them were male students, and 62 were female students. In terms of their educational background, 36 students came from secondary

schools where Chinese was the medium of instruction, while 39 students graduated from secondary schools where English was the medium. There were also 20 students who received their secondary education in international schools, and 5 students did not disclose information about their prior educational background.

3.2. Rubric and marking procedures

The rubric used in this study remained the same one employed by teachers in the English for Humanities and Social Sciences course to maintain high ecological validity. It is an analytic rubric structured into three domains, including responsiveness to the given prompt, rhetorical appropriateness and effectiveness, and language use. Teachers award individual marks to each domain, aggregate the scores, and then convert the overall marks into a letter grade following the English department's conversion standards. The rubric consists of five bands: A, B, C, D, and F. However, the study only included essays graded within the A, B, and C range as these grades were more representative of overall student performance; Grades of D and F were historically rare. See Appendix A for more details about A, B, and C grade descriptions.

In terms of the writing assessment, all the narrative essays were marked by experienced English teachers from the Language Center of this university. They all have at least five-year work experience in tertiary-level education in HK. Prior to marking the assessments, all teachers undergo rigorous training procedures to ensure the assessment quality. First, the teachers need to participate in a mandatory standardization session to calibrate their marking, following the rundown of the following activities: a) three benchmark samples that represent high (A level), middle (B level) and low (C level) writing quality are provided for the teachers to review; b) three additional samples are then assessed based on the assignment rubric in the standardization session by individual teachers; c) the course coordinator leads a discussion on the marks and addresses any questions or concerns regarding specific assessment samples; d) the final scores and grades of the assessment samples are announced to ensure a uniform understanding of the marking criteria. After all teachers mark student assignments, they need to submit their marksheet along with three samples representative of the three ranges (i.e., high, mid, and low) to the course team, which usually consists of the course coordinator and four experienced teachers. Upon collecting all the teachers' moderation materials and samples, the course team holds a moderation session for review. If the course team does not agree on any samples or notices any

course sections having abnormal cases (e.g., high average, narrow distribution), the corresponding teachers will be contacted for justification or re-adjustments.

3.3. Chatbot building and application

The GenAI tool used for the present study is named Poe AI (http://poe.com/). It was selected due to its wide accessibility in HK compared to ChatGPT launched by Open AI. Nevertheless, similar to ChatGPT, Poe allows users to build their customized Chatbots. In this study, we chose the GPT series (i.e., GPT-4o) as our base bot, considering its model was most recently updated in November 2024 among all the other options. Then, we provided seed information in the Prompt section, instructing the Chatbot how to behave and respond to user messages. As the goal of building this Chatbot is to resemble the professional knowledge of a human teacher in the English for Humanities and Social Sciences course, we provided various types of backgrounds, including 1) the description of the Chatbot's role; 2) students' linguistic and educational backgrounds; 3) the marking procedures and its criteria; 4) other cautionary notes. In addition, the Chatbot configuration includes a Knowledge Base where users can upload additional training data to improve its output accuracy. In total, we uploaded four sets of materials, which are 1) the assessment rubric with detailed descriptors for each grade level; 2) three sets of benchmark samples used in the actual teacher standardization; 3) the course book; and 4) the assessment instructions for students (See Appendix B for the chatbot configuration). By doing this, the customized Chatbot is trained with the same procedure that the teachers need to follow before marking students' narrative essays. To the best of our knowledge, most chatbots used for AI assessment, except for the one in Shin and Lee (2024), were generic. They were not trained with extensive materials designed to ensure AI possessed the same fundamental knowledge as a human teacher necessitated to conduct an assessment.

The three steps were applied to use the Chatbot to score the 100-timed narrative essays. First, the essays were all converted to PDF format for consistency to avoid any format change (e.g., paragraphing, spacing) that might influence the Chatbot's scoring. Second, the GenAI prompt was developed to conduct the automated scoring task. We adopted the zero-shot prompting approach when starting the conversation with the customized AI Chatbot. In other words, no further output examples were given in addition to the ones that had been uploaded in the Knowledge Base section. Before marking each essay, we asked the Chatbot to strictly follow

the sequence of actions: 1) carefully read the assessment-related PDFs uploaded in the Knowledge section; 2) generate marks including overall marks and analytical marks for specific assessment domains; 3) offer brief justifications for its scoring results (See Appendix C for the example prompt used to instruct the customized AI Chatbot to conduct scoring). Third, all marks and justification notes from the Chatbot were manually copied and pasted into our dataset. To prevent any intervention from previous marking judgments on the new round of scoring, we also initiated a new chat with the Chatbot each time, ensuring there was no presence of prior conversational history. Before actual implementation, we piloted our Chatbot to ensure it could understand the uploaded files and score the essays based on the assessment rubric. The instructions were fine-tuned multiple times until the Chatbot achieved stable output.

3.4. Data analysis

First, Cohen's Weighted Kappa statistic was employed to indicate the level of agreement between the two grading sources, assessing the reliability of the grading provided by the Chatbot in comparison to that of English teachers. The Weighted Kappa value was selected for the following two reasons: 1) the grade levels are ordinal variables (A > B > C), and 2) some disagreements are considered more severe than others (e.g., the difference between A and C is greater than that of A and B). Next, the correlations between the overall scores given by the Chatbot and by the English teachers were examined using Pearson correlation tests to evaluate the Chatbot's performance in writing assessment. The majority of the marks had shown a normal distribution, according to the QQ plots generated using the Statistical Package for Social Sciences (SPSS), as shown in Appendix D.

It should be noted that the 100 narrative essays did not encompass two specific types of cases, both of which could potentially influence the results of the correlation analysis.

These exclusions consisted of a) essays identified as containing instances of plagiarism and b) essays that received additional score deductions from the English teachers due to factors such as late submission. Consequently, all the narrative essays fell into three grade ranges: A (excellent), B (good), and C (average), and there was no essay with D and F.

4. Results

4.1. Descriptive statistics

This section includes descriptive statistics of the scores and grades assigned by the Chatbot and the English teachers. Additionally, the results of the level of agreement and correlation analysis are illustrated. Figure 1 demonstrates the descriptive statistics (i.e., mean and standard deviation) of the total scores assigned by the Chatbot and the English teachers for the 100 narrative essays. The total score was 35 points: The mean score from the Chatbot (27.08, which fell into the B+ range) was slightly higher than the mean score from the English teachers (25.13, which fell into the B range). As shown by the standard deviations, the final overall scores from the English teachers had a greater distribution (1.91) than the scores from the Chatbot (1.05). In other words, this indicates that compared to the English teachers, the Chatbot gave the scores at a narrower distribution with a relatively higher mean. More details about the frequencies of scores assigned are provided in the histograms in Appendix E.

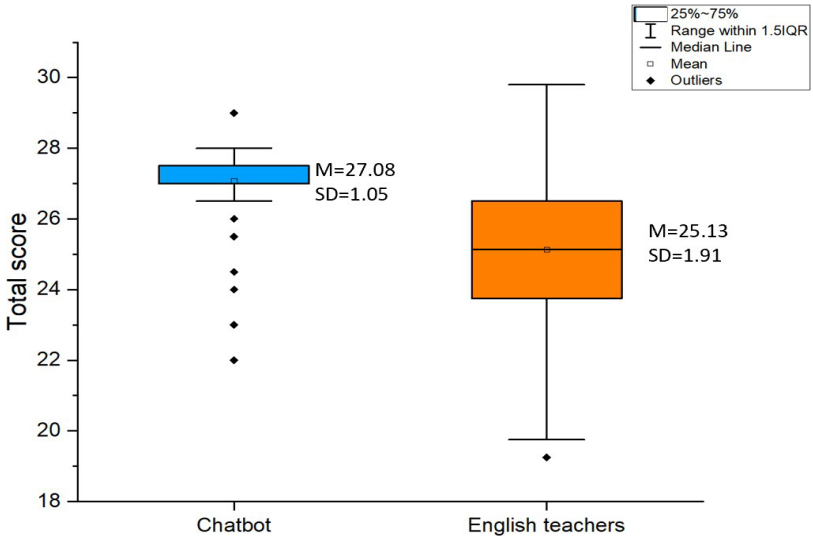


Figure 1. Scores assigned by the Chatbot and the English teachers

Table 1. Grades assigned by the Chatbot and the English teachers

		English teachers			Total
		A	B	C	
Chatbot	A	1	7	0	8
	B	4	81	6	91
	C	0	0	1	1
Total		5	88	7	100

Table 1 presents the grades assigned by the Chatbot and the English teachers to the narrative essays. The majority of the essays were graded within the B range by both the Chatbot (n = 91) and the English teachers (n = 88). However, notable differences were observed in the distribution of grades in other ranges. The Chatbot graded a higher number of essays (n = 8) in the A range than the English teachers (5). Conversely, the English teachers awarded more essays (n = 7) a grade in the C range compared to the Chatbot (n = 1). These findings suggest variations in grading tendencies between the Chatbot and the human evaluators. In general, it is not out of our expectation that more grades should fall into the B range due to the normal distribution nature of writing quality.

4.2. Level of Agreement and Correlational Analysis

To answer Research Question 1, the level of agreement between the grade levels (i.e., A, B and C) assigned by the Chatbot and the English teachers was assessed using Cohen's Weighted Kappa coefficient. As shown in Table 2, the analysis revealed a slight level of agreement (Weighted Kappa value of $\kappa = 0.153$, $N = 100$) between the Chatbot and the human raters according to the guidelines of Landis and Koch (1977). This indicates that the consistency in grading was weak, and the grading patterns of the Chatbot and the English teachers were not strongly aligned. However, the Kappa coefficient was significantly different from zero ($p = 0.017$), meaning the agreement between the Chatbot and the human raters was not due to random chance. More specifically, the level of agreement between the Chatbot and the human raters across nine specific grade levels (from A+ to C-) was poor but significant (Weighted Kappa value of $\kappa = -0.082$, p

= 0.013, N = 100), which indicates limited alignment in the grading of narrative essays.

Table 2. Level of agreement between grades assigned by the Chatbot and the English teachers

Grade levels	Chatbot – English teachers						
	Weighted Kappa	Asymptotic			95% Asymptotic Confidence Interval		N
		Std. Error	z	Sig.	Lower Bound	Upper Bound	
General (A, B and C)	0.153	0.117	2.376	0.017	-0.077	0.382	100
Specific (from A+ to C-)	0.082	0.039	2.473	0.013	0.004	0.159	100

Table 3. Correlation between the total scores assigned by the Chatbot and the English teachers

	Total scores by the English teachers	
Total scores by the Chatbot	Pearson Correlation (r)	0.446**
	Sig. (2-tailed)	<0.001
	N	100

*** Correlation is significant at the 0.01 level (2-tailed).*

To answer Research Question 2, a Pearson correlation test was computed to investigate how the total scores given by the Chatbot are correlated with the scores from the English teachers. Table 3 reveals that in terms of the significance of the correlation, the p-value ($p < 0.001$) indicates that there was a significant correlation between the total scores given by the Chatbot and the English teachers. Additionally, the correlation

coefficient ($r = 0.446$) presented a positive and moderate correlation.

5. Discussion

The findings suggest that ChatGPT-driven customized Chatbots may still be premature as an independent AES tool to be implemented in classroom or exam settings. According to descriptive statistics in this study, the mean score of the Chatbot was higher than that of the teacher scores with a narrower scoring distribution. This illustrates the Chatbot's leniency in marking, aligning with the findings from Tate et al. (2024) and Shin and Lee (2024). One possible explanation for the narrow scoring distribution given by the Chatbot is its inability to discern the nuanced difference between the students' essays when multiple grade levels share the same descriptors (e.g., A+, A, A-) on a rubric. Unlike the Chatbot with limited exposure to standardization samples, teachers tend to have natural instincts about the boundaries of each level after years of rigorous training and marking experience.

Furthermore, the analysis revealed a slight agreement between the teachers' scores and the Chatbot-generated scores in broad ranges (i.e., A, B, C). The agreement was much weaker when specific grade levels were taken into consideration (i.e., A+, A, A-, B+, B, B-, C+, C, C-). This result is not that surprising. Compared to Mizumoto and Eguchi (2023), they reported 54.33% of the exact agreement, but they used TOEFL iBT writing tasks (a high- stake test) that were produced in a more rigorously controlled setting. Our result is more consistent with Shabara et al. (2024), showing weak inter-rater reliability between their total scores when a more complex 0-100 point analytic rubric was employed. However, it is worth mentioning that Mizumoto and Eguchi (2023) acknowledged the value of GenAI, as the scoring outcomes were generally reflective of the three proficiency levels in the TOEFL iBT (i.e., low, mid, and high). This finding contradicts the observation in our study. The difference may be attributed to three factors. First, the rubric types are different. The assessment rubric with the maximum range of 1-15 points for a specific assessment domain in our study may adversely affect the GenAI's reasoning performance. This issue arises particularly because multiple score ranges within the same grade level share identical descriptors. The similar concern is also raised by Shabara et al. (2024).

Second, GenAI tools such as ChatGPT may be more familiar with the scoring criteria for certain types of academic genres (e.g., argumentative

essays). According to statistics published on the IELTS official website, more than 4 million IELTS exams were conducted in 2023 (IELTS, 2024). Another international language proficiency test, TOEFL iBT, is equally popular driven by the growing trend of overseas study. Although past exam results are likely inaccessible for ChatGPT to be used as training data due to confidentiality, there is an observable pattern that language institutions and students have been utilizing GenAI tools to support their language instruction which includes developing instructional materials and administering mock exams. These teaching and learning activities conducted through ChatGPT are valuable sources of training and feedback to augment its processing capability and output accuracy for genres such as argumentative essays. However, the narrative essay in our study has its own genre characteristics that ChatGPT might not developed sufficient knowledge to identify, thereby influencing its decision to automatically assess students' writing with precision. Third, the weighting of scores on each specific rubric domain may also influence the reliability of the writing scores. Numerous studies have reported that the strength of AES tools, including GenAI, lies in scoring linguistic features such as vocabulary and syntax rather than content, development, and creativity (Dikli, 2006; Wang & Brown, 2008; Shabara et al., 2024). Shabara (2024) also found a weak level of agreement on domains such as content and organization, causing the overall poor inter-rater reliability between the writing scores. Although the breakdown of scores on each domain is beyond this study's scope of discussion, we hypothesize that the greater weighting of content and organization (20 out of 35 marks) over language (15 out of 35 marks) in our rubric design may have further highlighted GenAI's weakness.

Nevertheless, this study presented a positive and moderate correlation between the teachers' and the Chatbot's total scores (r=0.446). This finding has marked significant progress in the performance of GenAI tools because the strength of correlation in our study surpasses most others except for only two studies (i.e., Kooli & Yusuf, 2024; Shin & Lee, 2024). Kooli and Yusuf (2024) identified a moderate correlation ($\rho = 0.450$) between the holistic scores by ChatGPT 3.5 and those by the human raters on 25 short responses completed under an exam setting with an average length of 250 words. Shin and Lee (2024) reported an even stronger correlation on all assessment domains on 50 essays of 80-120 words using the ChatGPT 4-based Chatbot (r= .91 for Organization, r= .93 for Language Use, r= .95 for Task Completion and Content). The correlation in our study is comparable to the correlation in Kooli and Yusuf (2024), but in contrast to Shin and Lee (2024), the correlation is much weaker. The contrast can

be caused by multiple factors, but the most salient two are sample size and length requirements: a) the sample size of our narrative essays was 100, which was significantly larger than Shin and Lee (2024), only 50 argumentative essays; 2) university students in our study produced longer texts, averaging approximately 600 to 800 words, whereas secondary schools students only produced 80 to 120 words in Shin and Lee (2024). Thus, a customized Chatbot with prior training has the potential to be applied for automated writing assessment for narrative essays, due to its moderate correlation with teacher scores, but we need to be aware of the slight level of agreement. There is still a large room for improvement.

6. Conclusions and Implications

From the perspective of the level of agreement and correlation, the study explored the relationship between GenAI scores and teacher scores on narrative essays in a compulsory English course at a HK university. Through correlation analyses, our study indicates that there is a positive and moderate correlation between the overall scores. This finding is attributed to the careful design of a customized AI Chatbot that emulates the exact marking procedures that teachers need to go through. With exposure to customized training data, we see an opportunity for GenAI as an AES tool to be improved. However, the analyses on reliability are less than ideal, showing only a slight level of agreement between the GenAI scores and teacher scores. Three main reasons have been summarized for weak reliability, including the clarity of rubric descriptors in each grade level, imbalanced weightings between content, organization, and language, and GenAI's relatively weak familiarity with the new type of genre.

The study is not without limitations. First, similar to other research studies, we must admit that one of the main challenges that cannot be handled is the potential internal inconsistency of scoring judgments among GenAI tools. This inconsistency arises mainly because GenAI is inherently characterized by randomness, meaning that different outputs can be generated when the same prompt is used (Schade, 2023). Meanwhile, GenAI is constantly evolving, and model updates between scoring periods could lead to variations in scoring outcomes. Second, this study exclusively focused on quantitative analyses. While justifications for each round of scoring were generated from GenAI in this study, this approach was intended to avoid any potential AI hallucination issues. Further investigations can be conducted to analyze qualitative feedback given by both teachers and GenAI in order to fully grasp the causes of scoring discrepancies. Third, we acknowledge that this is a small- scale study with

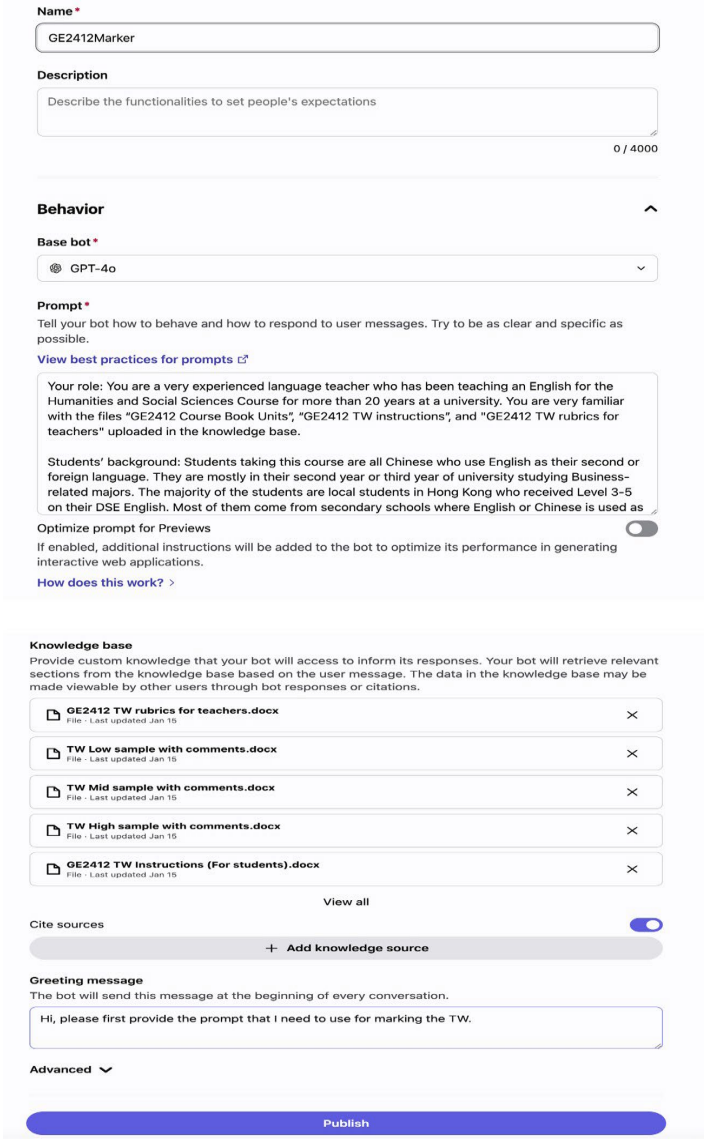
a focus on analyses of the alignment of the total scores between humans and GenAI; in the future, we hope to conduct a follow-up investigation that incorporates the dimension of analytical scores on each specific assessment domain. This additional information can contribute to a more holistic understanding of the strengths and weaknesses of GenAI's scoring ability.

The study also presents some implications. First, the study exposes both the strengths and weaknesses of using GenAI as an AES tool. It suggests directions to address the previously identified limitations if a customized Chatbot is implemented for assessment purposes. Additionally, the customized Chatbot used in this study and its potential future updates can be shared with teachers via a link in the English for Humanities and Social Sciences course. Teachers can have direct access to the Chatbot for use through the link upon their request without any changes being made, allowing them to experiment with it in different phases of a writing classroom, such as instruction and formative assessment. Second, despite an enthusiastic call in many studies for substituting teachers with AI-assisted raters in assessment duties, our study steers teachers toward the development of a fair and rational understanding of GenAI's ability. Human judgment, at this point, still need to play an essential role in writing evaluations. However, to expedite the review process, teachers are recommended to use GenAI to support them in providing formative feedback in essay drafts produced by students. This can greatly save tremendous time and energy in many English courses that are in favor of the process-oriented writing approach. As the application of GenAI will continuously penetrate the education realm, it is important for teachers to integrate GenAI as part of their writing instruction, thereby preparing students to become more digitally literate. For example, in an English writing course, teachers can present students with both human evaluations and AI-generated evaluations of sample essays, creating opportunities for discussion about what GenAI can or cannot achieve and how these evaluations can inform their writing process. Having an open discussion can educate students to use AI technologies more responsibly and ethically.

Appendix A

Grade levels	Descriptors
A	● The essay presents a comprehensive response to the prompt with a clear focus, well supported by evidence in an appropriate level of detail. ● The essay is organised effectively and conforms fully to the expectations of the genre and register. ● The text's lexicogrammar is extremely accurate and idiomatic.
B	● The essay is responsive to the prompt, focused, and generally well supported by evidence in sufficient detail. ● The essay is generally organised effectively and conforms to the expectations of the genre and register with only minor deviations. ● The text's lexicogrammar is generally accurate and idiomatic with only minor errors which do not negatively affect comprehension.
C	● The essay is generally responsive to the prompt although there may be some lack of focus, and evidence and detail may be lacking to some extent. ● The essay has a good organisational structure and broadly conforms to the expectations of the genre and register with some exceptions. ● The text's lexicogrammar is generally accurate and idiomatic with relatively few errors which negatively affect comprehension to a limited extent.

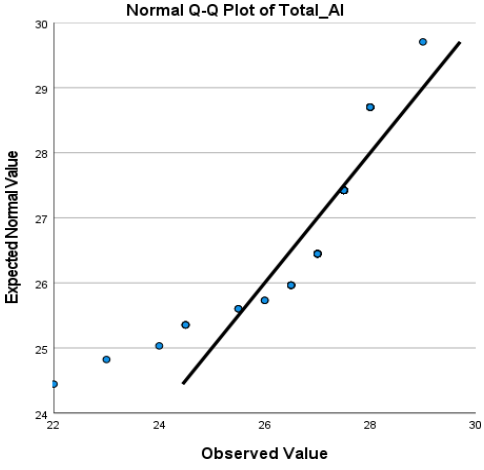
Appendix B


Name *
GE2412Marker
Description
Describe the functionalities to set people's expectations
0 / 4000
Behavior
Base bot *
GPT-4o
Prompt *
Tell your bot how to behave and how to respond to user messages. Try to be as clear and specific as possible.
View best practices for prompts
Your role: You are a very experienced language teacher who has been teaching an English for the Humanities and Social Sciences Course for more than 20 years at a university. You are very familiar with the files "GE2412 Course Book Units", "GE2412 TW instructions", and "GE2412 TW rubrics for teachers" uploaded in the knowledge base.
Students' background: Students taking this course are all Chinese who use English as their second or foreign language. They are mostly in their second year or third year of university studying Business-related majors. The majority of the students are local students in Hong Kong who received Level 3-5 on their DSE English. Most of them come from secondary schools where English or Chinese is used as
Optimize prompt for Previews
If enabled, additional instructions will be added to the bot to optimize its performance in generating interactive web applications.
How does this work?
Knowledge base
Provide custom knowledge that your bot will access to inform its responses. Your bot will retrieve relevant sections from the knowledge base based on the user message. The data in the knowledge base may be made viewable by other users through bot responses or citations.
GE2412 TW rubrics for teachers.docx
File · Last updated Jan 15
TW Low sample with comments.docx
File · Last updated Jan 15
TW Mid sample with comments.docx
File · Last updated Jan 15
TW High sample with comments.docx
File · Last updated Jan 15
GE2412 TW Instructions (For students).docx
File · Last updated Jan 15
View all
Cite sources
+ Add knowledge source
Greeting message
The bot will send this message at the beginning of every conversation.
Hi, please first provide the prompt that I need to use for marking the TW.
Advanced
Publish

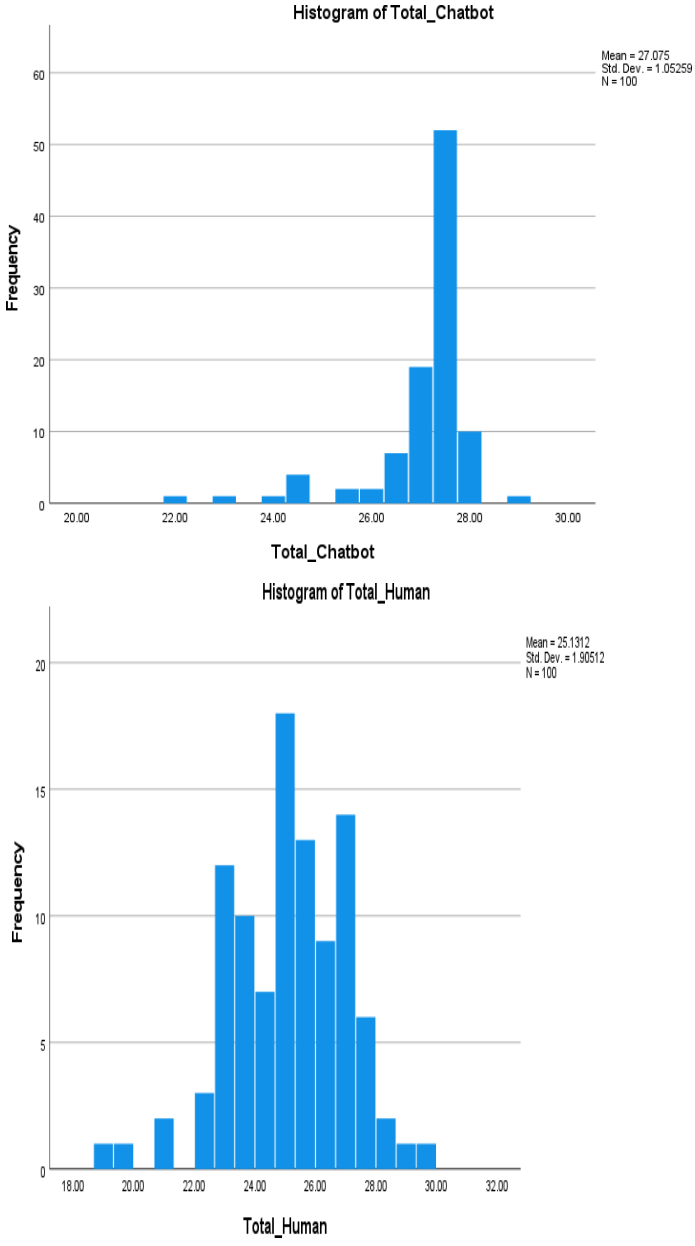
Appendix C

Mark the new PDF file in the attachment based on this prompt. [Insert the prompt that students responded to in the actual exam]. Output the results in a table format including the file ID, individual marks and grades, and the final marks (out of 35 points). In addition to the marks and grades, provide a short justification for your analytical and total marks.

Appendix D



Appendix E


Histogram of Total_Chatbot
Mean = 27.075
Std. Dev. = 1.05259
N = 100
Frequency
Total_Chatbot
Histogram of Total_Human
Mean = 25.1312
Std. Dev. = 1.90512
N = 100
Frequency
Total_Human

References

Bui, N. M., and J. S. Barrot. "ChatGPT as an Automated Essay Scoring Tool in the Writing Classrooms: How It Compares with Human Scoring." *Education and Information Technologies*, 2024, pp. 1–18, https://link.springer.com/article/10.1007/s10639-024-12891-w.

Dikli, S. "An Overview of Automated Scoring of Essays." *The Journal of Technology, Learning and Assessment*, vol. 5, no. 1, 2006, https://ejournals.bc.edu/index.php/jtla/article/view/1640.

Geçkin, V., et al. "Assessing Second-Language Academic Writing: AI vs. Human Raters." *Journal of Educational Technology and Online Learning*, vol. 6, no. 4, 2023, pp. 1096–1108, https://dergipark.org.tr/en/pub/jetol/issue/80405/1336599.

Guo, K., and D. Wang. "To Resist It or to Embrace It? Examining ChatGPT's Potential to Support Teacher Feedback in EFL Writing." *Education and Information Technologies*, vol. 29, no. 7, 2023, pp. 8435–63, https://link.springer.com/article/10.1007/s10639-023-12146-0.

Guo, K., et al. "Effects of an AI-Supported Approach to Peer Feedback on University EFL Students' Feedback Quality and Writing Ability." *The Internet and Higher Education*, vol. 63, 2024, article 100962, https://doi.org/10.1016/j.iheduc.2024.100962.

"IELTS Continues to Lead, Support, and Empower around the World." IELTS, 11 Jan. 2024, https://ielts.org/news-and-insights/ielts-trusted-by-millions-in-2023.

Kim, J., et al. "Exploring Students' Perspectives on Generative AI-Assisted Academic Writing." *Education and Information Technologies*, 2024, pp. 1–36, https://link.springer.com/article/10.1007/s10639-024-12878-7.

Kooli, C., and N. Yusuf. "Transforming Educational Assessment: Insights into the Use of ChatGPT and Large Language Models in Grading." *International Journal of Human–Computer Interaction*, 2024, pp. 1–12, https://www.tandfonline.com/doi/full/10.1080/10447318.2024.2338330.

Labov, W. *Language in the Inner City*. U of Pennsylvania P, 1972.

McAdams, D. P. "American Identity: The Redemptive Self." *The General Psychologist*, vol. 43, no. 1, 2008, pp. 20–27,

https://www.scribd.com/document/348771837/American-identity-the-redemptive-self-pdf.

Mizumoto, A., and M. Eguchi. "Exploring the Potential of Using an AI Language Model for Automated Essay Scoring." *Research Methods in Applied Linguistics*, vol. 2, no. 2, 2023, article 100050, https://www.sciencedirect.com/science/article/pii/S2772766123000101.

Özyıldırım, I. "Narrative Analysis: An Analysis of Oral and Written Strategies in Personal Experience Narratives." *Journal of Pragmatics*, vol. 41, no. 6, 2009, pp. 1209–22, https://www.sciencedirect.com/science/article/abs/pii/S037821660900023X.

Palaña, L. S. *Narrative Stylistics*. 2014, https://www.slideshare.net/lordinnisia/narrative-stylistics.

Schade, M. "How ChatGPT and Our Language Models Are Developed." OpenAI Help Center, 28 Oct. 2023, https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed.

Shabara, R., et al. "Teachers or ChatGPT: The Issue of Accuracy and Consistency in L2 Assessment." *Teaching English with Technology*, vol. 24, no. 2, 2024, pp. 71–92, https://www.ivysci.com/en/articles/4180224.

Shin, D., and Jang Ho Lee. "Exploratory Study on the Potential of ChatGPT as a Rater of Second Language Writing." *Education and Information Technologies*, 2024, pp. 1–23, https://doi.org/10.1007/s10639-024-12817-6.

Su, Y., et al. "Collaborating with ChatGPT in Argumentative Writing Classrooms." *Assessing Writing*, vol. 57, 2023, article 100752, https://www.sciencedirect.com/science/article/abs/pii/S1075293523000600.

Tate, T. P., et al. "Can AI Provide Useful Holistic Essay Scoring?" *Computers and Education: Artificial Intelligence*, vol. 7, 2024, article 100255, https://www.sciencedirect.com/science/article/pii/S2666920X24000584.

Teng, M. F. "'ChatGPT Is the Companion, Not Enemies': EFL Learners' Perceptions and Experiences in Using ChatGPT for Feedback in Writing." *Computers and Education: Artificial Intelligence*, vol. 7, 2024, article 100270,

https://www.sciencedirect.com/science/article/pii/S2666920X24000730.

Wang, J., and M. S. Brown. "Automated Essay Scoring versus Human Scoring: A Correlational Study." *Contemporary Issues in Technology and Teacher Education*, vol. 8, no. 4, 2008, pp. 310–25, https://www.learntechlib.org/p/25295/.

Yamashita, T. "An Application of Many-Facet Rasch Measurement to Evaluate Automated Essay Scoring: A Case of ChatGPT-4.0." *Research Methods in Applied Linguistics*, vol. 3, no. 3, 2024, article 100133, https://doi.org/10.1016/j.rmal.2024.100133.

Yancey, K. P., et al. "Rating Short L2 Essays on the CEFR Scale with GPT-4." Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), July 2023, pp. 576–84, https://aclanthology.org/2023.bea-1.49/.

Zhang, R., et al. "Chatbot-Based Learning of Logical Fallacies in EFL Writing: Perceived Effectiveness in Improving Target Knowledge and Learner Motivation." *Interactive Learning Environments*, vol. 32, no. 9, 2024, pp. 5552–69, https://www.tandfonline.com/doi/abs/10.1080/10494820.2023.2220374.

Address for Correspondence

Ge Lan
Department of English
City University of Hong Kong
18 Tat Hong Ave.
Kowloon Tong, Hong Kong

gelan4@cityu.edu.hk

Jie Yang
Department of English
City University of Hong Kong
18 Tat Hong Ave.
Kowloon Tong, Hong Kong

jyang399@cityu.edu.hk

Xuan-Zi He
Department of English
City University of Hong Kong
18 Tat Hong Ave.
Kowloon Tong, Hong Kong

xuanzihe2-c@my.cityu.edu.hk

Yi Li
Chan Feng Men-ling Chan Shuk-lin Language Centre
City University of Hong Kong
Tat Chee Ave.
Kowloon Tong, Hong Kong

yili252@cityu.edu.hk

Submitted Date: February 25, 2025
Accepted Date: April 22, 2025