Discrepancies Between ChatGPT and Vietnamese EFL Teachers in Writing Assessment

Ho Nhut Nam[1*], Tra Thi Cam Thu[1], Pham Thi Quyen[1], Tran Phuong Thanh[1]

[1] Faculty of Foreign Languages, Industrial University of Ho Chi Minh City, Vietnam
[*]Corresponding author's email: honhutnam.2505@gmail.com
[*] https://orcid.org/0009-0007-8586-9555
https://doi.org/10.54855/ijaile.26311

Received: 26/08/2025 Revision: 13/03/2026 Accepted: 25/03/2026 Online: 29/04/2026

ABSTRACT

Artificial intelligence, particularly ChatGPT, strongly influences education, especially in modern EFL classrooms. However, a significant gap still exists in Vietnam regarding how ChatGPT's writing assessments differ from those of EFL teachers. To address this gap, this study directly compares ChatGPT's writing assessments with those of Vietnamese EFL teachers to identify key discrepancies. Twenty experienced teachers working at English centers and universities in Ho Chi Minh City participated in this quantitative research. ChatGPT and human raters used an analytic rubric adapted from a standardized rubric to assess the written works of twenty students. A paired-sample t-test was then employed to compare the scores of human raters and ChatGPT in assessing twenty students' essays across the rubric criteria. The findings revealed a statistically significant difference between ChatGPT and Vietnamese EFL teachers in writing evaluations ($p = .034$), with ChatGPT (M = 16.08, SD = 3.29) assigning a higher total score than human raters (M = 14.75, SD = 2.96). Particularly, ChatGPT tended to give higher scores than human raters across all criteria, including content, organization, vocabulary, and grammar, with the content criterion showing the greatest discrepancy.

Keywords: ChatGPT, Vietnamese EFL teacher, writing, discrepancy, rubric

Introduction

Background of the study

In this technological era, it is uncommon for a new development to make waves as quickly as ChatGPT has within just a few months (Hong, 2023). This artificial intelligence (AI)-assisted model has a significant impact on language education. ChatGPT has gained popularity and proven its ability not only to provide instant assistance and personalized adaptive learning for learners but also to help educators modernize their teaching approaches and assessment methods, thereby improving teachers' work efficiency and saving time (Rudolph et al., 2023; Hoang et al., 2025). Along with the emergence of ChatGPT, automated scoring systems have amazed linguistic communities with their ability to provide instant feedback and maintain consistency during the grading process. These systems help reduce the burden of manual

CITATION | Ho, N. N., Tra, T. C. T., Pham, T. Q., & Tran, P. T. (2026). Discrepancies Between ChatGPT and Vietnamese EFL Teachers in Writing Assessment. *International Journal of AI in Language Education, 3*(1), 1-16. DOI: https://doi.org/10.54855/ijaile.26311

grading, potential bias, and subjectivity introduced by human raters (Ludwig et al., 2021), and EFL teachers generally hold a positive outlook on employing ChatGPT as a supportive tool in writing classes (Nguyen, 2023).

Statement of the problems and the purpose of the study

It is obvious that grading students' writing papers takes a lot of time for EFL teachers (Baker, 2014), so educators around the world have employed ChatGPT as a writing assessment tool to reduce teachers' workload. Because of this, many studies have been conducted to find out the effectiveness of ChatGPT in grading writing papers. According to Bucol and Sangkawong (2024), although ChatGPT has proven its ability to efficiently conduct writing assessments using customized rubrics, it still has certain disadvantages that require human observation to enhance evaluation accuracy. In contrast, Li et al. (2024) found that ChatGPT had a greater reliability coefficient than human raters. These mixed findings show that using ChatGPT as a writing assessment tool remains a topic of debate regarding its accuracy and consistency compared to human evaluators.

The same pattern can be seen in Vietnam's EFL settings. EFL teachers have to assess a wide range of written work, and this workload puts significant pressure on their teaching effectiveness (Hang, 2021). To address this challenge, many institutions, such as universities and schools, integrate ChatGPT across many fields of education, especially in foreign language teaching (Pham et al., 2025), because it has received numerous accolades as a diligent assistant from English educators (Li et al., 2024). However, teachers also voice many concerns about ChatGPT's overreliance, integrity, and the need to integrate it properly in their teaching process (Vo & Huynh, 2025). As a result, many comparisons of the accuracy of writing assessment between human raters and ChatGPT have been conducted. Poláková et al. (2024) concluded that ChatGPT's evaluative ability is still not as good as human teachers' insight in accurately judging written works, based on the fact that ChatGPT cannot replace the unique ability of teachers to apply pedagogical judgment and contextual understanding in evaluating students' writing (Atasoy et al., 2025). It is noticeable that concerns about scoring accuracy and the discrepancy in writing evaluation between ChatGPT and human raters are now widely recognized worldwide, including in Vietnam. Although previous studies have examined the use of ChatGPT in writing assessments compared with human raters, most have been conducted outside Vietnam. Consequently, empirical evidence comparing ChatGPT and Vietnamese EFL teachers' evaluations of students' writing remains limited. Therefore, the primary objective of this research is to examine whether evaluations of students' writing using an analytical rubric by Vietnamese EFL teachers differ from those conducted by ChatGPT, and to find out which criterion shows the greatest discrepancy

The significance of the study

This study is significant because it examines ChatGPT, one of the most modern tools of the 21st century, as a means to assess students' written work in Vietnamese EFL classrooms. Along with that, this paper compares ChatGPT writing assessments with Vietnamese EFL raters using an analytic rubric so that it helps to find out which criteria in an analytic rubric show the most discrepancy. Therefore, the findings may not only offer insights into how AI can effectively support teachers and reduce their grading workload but also encourage the effective and accurate integration of ChatGPT and AI in general in assessing writing in educational settings.

Literature review

Traditionally, most writing assessments are conducted by human teachers. Despite using and developing a range of assessment techniques, human graders often lack confidence in their assessment knowledge (Berry et al., 2019). Thanks to technological advancements, AI-driven grading systems have been successfully introduced into language education and are now part of language assessment. In EFL, these AI-based grading systems have many advantages, such as increased effectiveness and reduced assessment workload (Palmer et al., 2002). However, integrating AI-based grading systems might pose challenges, especially when applied to writing assessment.

Empirical studies have mostly examined the differences between ChatGPT and human raters in writing assessments from two main perspectives. While some studies have focused on identifying overall differences between AI and human graders without analyzing each rubric criterion in detail, others have examined criterion-level discrepancies, such as content, organization, or grammar. To fully reflect those differences, this review is organized into two main sections.

General discrepancies between ChatGPT and human raters in writing assessment

Empirical studies have reported mixed findings regarding the general differences between ChatGPT and human raters. Many studies suggest that human raters provide more meaningful and context-sensitive evaluations than ChatGPT. For example, Nguyen and Tran (2023) conducted research at the University of Foreign Language Studies to examine the efficacy of ChatGPT's writing evaluation in Vietnam's L2 context. Ten essays produced by advanced EFL students were graded by ChatGPT based on VSTEP criteria for both feedback and assigning grades. A seasoned senior instructor with expertise in teaching advanced English writing subsequently reviewed these outcomes. To further explore the teacher's perspective on ChatGPT's reliability and effectiveness, the instructor participated in a semi-structured interview. The findings revealed that ChatGPT provided generic, repetitive comments of varying length, whereas teachers delivered accurate, contextually relevant evaluations. Similarly, Bucol and Sangkawong (2024) highlighted ChatGPT's limitations as a writing assessment tool compared to human lecturers. The research was conducted at a Thai university and combined quantitative and qualitative methods. Their findings showed that human assessors could understand contextual insights and identify errors, whereas ChatGPT struggled. In addition, the authors noted the AI's tendency to grade leniently, a critical point for teachers to consider. Generally, the findings of the above studies suggest that human evaluators can interpret students' writing beyond surface-level features.

However, other studies have reported a relatively strong alignment between humans and ChatGPT in writing evaluation. For instance, Koraishi (2024) conducted a study to examine the reliability of ChatGPT as a writing assessment tool for IELTS Writing Task 2. Fifty-five sample writings from real examinations were used in this study. Using a quantitative approach that involved calculating correlations and comparing means, the authors concluded that although some inconsistencies were observed, ChatGPT's scores showed strong alignment with human graders' scores. These contrasting findings indicate that ChatGPT's reliability in writing assessment may vary depending on the context and evaluation criteria.

Some research has revealed significant scoring differences between ChatGPT and human raters. To examine the relationship between ChatGPT and human scoring in 200 argumentative essays, Bui and Barrot (2025) adopted a cross-sectional quantitative design, in which every essay was rated twice by ChatGPT and experienced raters, using a shared writing rubric. According to

correlational research, ChatGPT's scores did not closely correspond to those of a skilled human rater, with correlations ranging from low to average. Additionally, after two scoring rounds, ChatGPT was unable to maintain its consistency. Similarly, in 2025, Uyar and Büyükahıska conducted a quantitative study. The authors examined the feasibility and effectiveness of using ChatGPT to evaluate essays from EFL learners. The researchers analyzed the essays of 10 B2-level students and compared the grades assigned by human raters with those assigned by ChatGPT using the Wilcoxon signed-rank and Spearman correlation tests. The findings revealed that while ChatGPT can be a useful tool, its scores sometimes align with human raters, but the AI consistently gave significantly lower scores than the human raters. Furthermore, to examine the reliability and efficiency of AI-generated scores and the causes of inconsistent assessments between ChatGPT and human graders, Kim et al. (2024) used a mixed-methods approach with an explanatory sequential design comprising two distinct stages. Six PhD candidates in the Applied Linguistics and Technology program and ChatGPT participated in this study, and graded 74 essays selected from the Fall 2016 EPT administration data. The findings demonstrated that ChatGPT tended to penalize writers for some aspects that human raters consider less important. In addition, Steiss et al. (2024) conducted a study in Southern California that focused on the differences between instructors' and ChatGPT's grading quality when assessing 200 high school students' writing essays. By applying a quantitative approach, specifically descriptive statistics and one-way Analysis of Variance (ANOVA), the study revealed that teachers' results were more precise and specific to individual students' writing than those delivered via ChatGPT.

Overall, previous studies in writing assessments from ChatGPT and human evaluators share both similarities and differences. While ChatGPT can approximate human scoring in some contexts, some inconsistencies still persist in the grading pattern. However, few empirical studies have examined differences between scores provided by human raters and ChatGPT in the Vietnamese EFL context. Therefore, further studies are needed to better understand and investigate the discrepancies between ChatGPT and human raters in the Vietnamese EFL context.

Discrepancies between ChatGPT and human raters in writing assessment across rubric criteria

A growing body of research has examined discrepancies between ChatGPT and human raters at the level of rubric criteria. Many studies have reported inconsistencies across different writing evaluation features, such as content, organization, and language use, which suggest that ChatGPT-based evaluation may struggle with certain aspects of writing assessment. For example, Jackaria et al. (2024) conducted a descriptive–comparative study in the Philippines that aimed to compare the scoring results of twenty-eight junior students' essays graded by ChatGPT and by three human raters using a validated rubric. For ChatGPT ratings, essays were encoded and input into ChatGPT 3.5 using prompts and the rubric. These essays were also rated independently by three human raters using the same scoring rubric. Using the intraclass correlation coefficient (ICC) and the Wilcoxon test, the results showed that ChatGPT's writing assessment lacked consistency in content, organization, and language, and that ChatGPT tended to assign slightly higher scores. Similar discrepancies were reported by Topuz et al. (2025). The researchers analyzed 210 essays from 35 undergraduate students using a five-criterion rubric that included ideas, organization, coherence, support, style, and mechanics. Essays were evaluated by two raters and twice by ChatGPT. The research results revealed that ChatGPT and human raters differed across all categories (ideas, organization and coherence, support, style, and mechanics).

Other studies indicate that ChatGPT's grading performance may vary depending on the rubric

criteria being evaluated. In a Chinese EFL setting, Yang (2024) conducted a study using an independent-samples t-test and Pearson's correlation to compare the average scores of 82 Chinese students' writing samples, evaluated by ChatGPT and human raters from the Written English Corpus. According to the findings, ChatGPT performed poorly when evaluating the organization of EFL compositions since it was more challenging for AI systems to judge organizational elements than language or content elements. Similarly, Poláková et al. (2024) reported ChatGPT's shortcomings, particularly in correctly assessing grammatical features, underscoring the need to use it with caution. Taken together, these studies suggest that ChatGPT may face challenges in applying analytical rubric criteria when assessing EFL writing.

Additional research has examined ChatGPT's evaluation of specific writing features, particularly discourse elements such as coherence and cohesion. Using a mixed-methods approach, Yoon et al. (2023) investigated the effectiveness of ChatGPT's feedback in assessing 50 essays written by 12th-grade English learners at a high school in the USA, focusing on coherence and cohesion. Although ChatGPT scored correctly according to the rubric criteria, the author of this study did not recommend the use of ChatGPT to support evaluating English language writers' coherence/cohesion. This suggests that while ChatGPT can follow the rubric guidelines, its evaluation may require human raters' oversight, especially when assessing coherence and cohesion features.

However, not all studies report substantial misalignment between ChatGPT and human raters. Geckin et al. (2023) examined the agreement between ChatGPT-3.5 and human raters in assessing second-language academic writing. The study involved 43 Turkish first-year college students who were advanced EFL learners. In addition, a paragraph writing task and a holistic rubric were used for assessment, with the rubric introduced to both ChatGPT and five trained human raters through clear instructions and a standardization session. Using a correlational design, the study found significant agreement between the scores assigned by two human raters and ChatGPT-3.5, with the AI tool rating sentences more accurately, while the human raters considered both sentence details and overall writing quality. Furthermore, research showed that variation in writing assessment may stem from differences among human raters. A case in point is that González et al. (2017) investigated the discrepancy in assessing the same sample writings among the raters. The study used a mixed-methods design and was conducted at a public university in Mexico. Data was obtained from the five written samples scored by the 11 EFL teachers. The results showed that when teachers used the same rubric and had similar backgrounds, their scores also differed. These findings indicate that discrepancies in writing assessment are not only due to AI limitations but also to human judgment.

The aforementioned studies highlight differences between human raters and ChatGPT in writing assessment and show inconsistencies in ChatGPT's assessment across various criteria, including organization, coherence, and grammatical nuances. However, little research has been conducted to find which criteria had the greatest misalignment with human raters. This highlights the need to conduct this research to identify which writing assessment criteria exhibit the most discrepancy between ChatGPT and human raters.

Research Questions

This study aimed to respond to the following research questions to accomplish the study's objectives:

1. To what extent do ChatGPT's writing assessments differ from those of Vietnamese EFL teachers?
2. In which writing assessment criteria do the greatest discrepancies between ChatGPT and

human raters occur?

Methods

Pedagogical Setting & Participants

The research was conducted with the participation of 20 English teachers from prestigious English centers and universities. They have over 3 years of English-teaching experience in Ho Chi Minh City, and all the teachers are familiar with analytical writing rubrics. Their role was to assess students' essays using the same rubric across four criteria: content, organization, grammar, and vocabulary. In this study, ChatGPT version GPT-4.0 was used to grade the same student essays using the same rubric as the human raters. The same input prompt for ChatGPT was used to ensure the alignment. The essays were written by twenty non-major English students who are second-year students at the Academy of Public Administration and Governance. The students' English proficiency was not formally measured; it was estimated by the homeroom teacher's observation to be around the A2 level. According to the school curriculum, students have been taught to write a short opinion essay as part of their writing course. For this study, students took the regular test of their English writing course; they were asked to write an essay of at least 120 words in twenty minutes. Then, twenty essays were randomly chosen for the study.

Design of the Study

This study employed a quantitative approach as the primary method to explore differences between Vietnamese EFL teachers and ChatGPT in writing assessment. An analytic rubric, including four criteria of the writing (content, organization, vocabulary, and grammar), was constructed and adapted from IELTS Writing Band Descriptors developed by the International English Language Testing System (IELTS, 2023). The IELTS rubric provides clear descriptors and can be easily adjusted for lower-proficiency students. First, the original IELTS scale, which ranges from 0 to 9, was simplified into a 0-5 scoring scale to make the scoring practical and manageable for human raters to apply consistently. Secondly, the band descriptors were linguistically simplified to suit A2-level learners. For example, many complex descriptors related to argument development and advanced language use were simplified into clearer and easier descriptions that focused on idea relevance, basic organization, appropriate language use, and grammatical accuracy. In addition, the original IELTS criteria (task response, coherence and cohesion, lexical resource, and grammatical range and accuracy) were retained and renamed as content, organization, vocabulary, and grammar to ensure raters could understand and apply them consistently in the context of this study. The rubric used a 0-5 point scale for each of the four criteria, and each score level included detailed descriptions for every criterion. Each criterion was scored out of a maximum of 5 points, with the 4 criteria totaling 20 points. Half-point scores such as 1.5, 2.5, etc., were also allowed to be assigned by human raters and ChatGPT.

This rubric also underwent pilot testing before collecting data. Five teachers with more than three years of teaching experience and evaluating writing were given a rubric and five opinion paragraphs. They were asked to evaluate five sample opinion paragraphs using the adapted rubric. After completing the scoring process, the five teachers were asked to provide written comments regarding the clarity of the scoring descriptors. During the assessment process, raters showed some confusion related to the descriptions of the content criterion. The teachers found it hard to distinguish between score levels, but no concern was raised about the other criteria. The collected feedback was used to clarify the descriptions of the content criterion, making it

easier for raters to apply.

Data collection & analysis

A sample of twenty essays with the topic "*Write an essay in at least 120 words for or against the statement: Sports should be taught at school*" was obtained from the Academy of Public Administration and Governance in Ho Chi Minh City (APAG). Students were assigned the writing task as part of their regular test; they had 20 minutes to complete it on paper, under careful observation from the homeroom teacher.

After randomly collecting students' written responses on paper, the researchers scanned the test papers and converted them into document files to facilitate the research. Each essay received two sets of scores across the four rubric criteria: one from a human rater and one from ChatGPT. The assessment process was divided into two stages. In the first stage, due to the teacher's workload and time constraints, each of the twenty raters was randomly assigned one essay to evaluate using the revised analytic rubric, and the raw scores were collected and entered into Google Sheets. The scores provided by 20 human raters were considered a collective representation of human evaluation to compare with ChatGPT's scores. In the second stage, the same 20 essays that had been rated by human raters were individually entered into ChatGPT for assessment using the same rubric as the one used by the human raters. In this stage, researchers uploaded students' essays to ChatGPT and asked it to transcribe them. Afterward, the researchers carefully reviewed the transcriptions to ensure they matched the students' handwritten essays. Following this, each essay was preceded by the prompt: "I will send you an essay written by a student. You have to give an accurate grade according to the given rubric. You are allowed to give half-point scores such as 0.5, 1.5, 2.5, 3.5, and 4.5. The topic for the essay is [topic]. The text is [the student's essay]. The same prompt was always used in each of the twenty essays to ensure consistency. Each essay was always assessed in the new "chat," which did not include any prior conversation to avoid ChatGPT's bias and ensure accuracy. The ChatGPT score was also collected and entered into Google Sheets. The teacher's score and ChatGPT's score were recorded in the same row, ensuring they corresponded to the same essay.

Next, the scores were analyzed using a paired-sample t-test in SPSS version 25. A paired-samples t-test was used to evaluate whether there was a difference between ChatGPT and human raters' scores. In this stage, the researchers compared the scores for each rubric element and the total score for each essay, graded by human raters and by ChatGPT. Additionally, to identify which criterion showed the greatest discrepancy between ChatGPT and human ratings, the researchers calculated the average score for each criterion from human raters and ChatGPT and compared them.

Results/Findings

Research question 1: To what extent do ChatGPT's writing assessments differ from those of Vietnamese EFL teachers?

Table 1

Total score by human raters and ChatGPT

	Mean	N	SD	Std. Error Mean
Total score by human raters	14.7500	20	2.97578	.66540
Total score by ChatGPT	16.0750	20	3.28984	.73563

	Mean	SD	Std. Error Mean	95% CI Lower	95% CI Upper	t	df	p
Total score by human raters - Total score by ChatGPT	-1.32500	2.58678	.57842	-2.53565	-.11435	-2.291	19	.034

Table 1 displays the results of a paired-samples t-test comparing total scores from human raters and ChatGPT. The statistics showed that the mean score assigned by human raters was lower (M=14.75, SD=2.98) than the mean score generated by ChatGPT (M=16.08, SD=3.29). The mean difference between the two sets of scores was -1.325, indicating that ChatGPT tends to give higher scores. The results revealed a statistically significant difference between the two scoring sources. $t(19) = -2.29$, $p = .034$, so the null hypothesis was rejected, suggesting the difference in scores was statistically significant. This indicates that ChatGPT's scoring of student writing was notably more generous than that of human raters.

Table 2

Content score by human raters and ChatGPT

	Mean	N	SD	Std. Error Mean
Content score by human raters	3.775	20	.6973	.1559
Content score by ChatGPT	4.275	20	.7860	.1758

	Mean	SD	Std. Error Mean	95% CI Lower	95% CI Upper	t	df	p
Content score by human raters - Content score by ChatGPT	-.5000	.7255	.1622	-.8395	-.1605	-3.082	19	.006

Table 2 presents the results of a paired-samples t-test comparing content scores from human raters (M = 3.78, SD = 0.70) and ChatGPT (M = 4.28, SD = 0.79). As shown in the table, the mean difference between the two sets of scores was -0.50 (SD = 0.73), indicating that ChatGPT tended to assign slightly higher scores to the content criterion than humans did. Moreover, the test included twenty essays that were separately graded by humans and by ChatGPT, so the degree of freedom (df) was (20)-1=19. The result revealed a significant difference ($p = .006$). Therefore, the alternative hypothesis was accepted, and the null hypothesis was rejected. It indicates a discrepancy between the scores given by human raters and by ChatGPT in content evaluation.

Table 3

Grammar score by human raters and ChatGPT

	Mean	N	SD	Std. Error Mean
Grammar score by human raters	3.525	20	.8807	.1969
Grammar score by ChatGPT	3.775	20	.8656	.1936

	Mean	SD	Std. Error Mean	95% CI		t	df	p
				Lower	Upper			
Grammar score by human raters - Grammar score by ChatGPT	-.2500	.8030	.1795	-.6258	-.1258	-1.392	19	.180

Table 3 presents the results of comparing scores between human raters and ChatGPT in the grammatical criteria. As shown in Table 3, the average grammar score given by human raters was 3.53 (M = 3.53, SD = 0.88), while ChatGPT's score was 3.78 (M = 3.78, SD = 0.87). It means that ChatGPT gave 0.25 points higher than human raters on average. Moreover, the paired-samples t-test revealed that the discrepancy in mean scores between ChatGPT and human raters on grammatical criteria (M = -0.25, SD = 0.80) was not statistically significant, $t(19) = -1.39$, $p = .180$. This suggests that the null hypothesis could not be rejected. Although ChatGPT tended to score higher than human ratings, the difference was not large enough to claim a real difference in ratings.

Table 4

Vocabulary score by human raters and ChatGPT

	Mean	N	SD	Std. Error Mean
Vocabulary score by human raters	3.750	20	1.0822	.2420
Vocabulary score by ChatGPT	3.875	20	.9159	.2048

	Mean	SD	Std. Error Mean	95% CI		t	df	p
				Lower	Upper			
Vocabulary score by human raters - Vocabulary score by ChatGPT	-.1250	.8565	.1915	-.5258	.2758	-.653	19	.522

From the statistical results of comparing the vocabulary scores between human raters and ChatGPT in Table 4, it can be observed that the mean vocabulary score given by human raters was slightly lower (M = 3.75, SD = 1.08) than that assigned by ChatGPT (M = 3.88, SD = 0.92). The test showed a mean difference of -0.13 between the two sets of scores, indicating that human raters scored slightly lower than ChatGPT. However, this difference was not statistically significant, $t(19) = -0.65$, $p = .522$. This means ChatGPT tended to assign slightly higher vocabulary scores, and there was no significant difference between the vocabulary scores assigned by human raters and those assigned by ChatGPT.

Table 5

Organization scores between human raters and ChatGPT

	Mean	N	SD	Std. Error Mean
Organization score by human raters	3.700	20	.8176	.1828
Organization score by ChatGPT	4.150	20	.8127	.1817

	Mean	SD	Std. Error Mean	95% CI		t	df	p
				Lower	Upper			
Organization score by human raters - Organization score by ChatGPT	-.4500	.7416	.1658	-.7971	-.1029	-2.714	19	.014

Table 5 demonstrates the results of analyzing the organization scores graded by human raters and by ChatGPT. It showed that the mean organization score assigned by human raters (M = 3.70, SD = 0.82) was lower than the mean given by ChatGPT (M = 4.15, SD = 0.81). A paired-sample t-test presented a mean difference of -0.45, showing that ChatGPT gave significantly higher scores than human raters. With $t(19) = -2.71$, $p = .014$. It is concluded that the difference between the two scoring sources was statistically significant. In short, ChatGPT is more lenient in evaluating organizational scores than human raters.

Research question 2: In which writing assessment criteria do the greatest discrepancies between ChatGPT and human raters occur?

Table 6

Differences in the mean scores between human raters and ChatGPT across four criteria.

Criterion	Mean scores of human raters	Mean scores of ChatGPT	Mean difference (Human - ChatGPT)
Content	3.78	4.28	-0.50
Organization	3.70	4.15	-0.45
Grammar	3.53	3.78	-0.25
Vocabulary	3.75	3.88	-0.13

Table 6 illustrates the differences in mean scores between human raters and ChatGPT. Regarding the content criterion, the mean score from human raters was 3.78, while ChatGPT assigned a higher score of 4.28, resulting in a mean difference of -0.50. It is clear that the mean difference for the content criterion was the largest among the four criteria; this substantial gap indicates ChatGPT's greater generosity in evaluating the content criterion compared to human raters. This discrepancy revealed differences in how human raters and ChatGPT interpret and prioritize content-related elements in students' writing. While human raters tend to be sensitive to subtle content-related issues, ChatGPT appears to focus on the overall quality of ideas, such as relevance and completeness.

In terms of the organization criterion, the mean scores from human raters and ChatGPT were

3.70 and 4.15, respectively, resulting in a mean difference of -0.45. This second noticeable discrepancy suggested that ChatGPT was inclined to tolerate organization-related errors that human raters penalize more strictly. These results also confirm that human raters might interpret the logical flow and coherence of ideas more critically. The gap between ChatGPT and human raters in the organization criterion was smaller than that in the content criterion, suggesting that ChatGPT and human raters might apply different perspectives when evaluating organizational aspects.

Moreover, the mean differences between human raters and ChatGPT on grammar and vocabulary criteria were relatively small, at -0.25 and -0.13, respectively. This indicated a higher level of alignment between human raters and ChatGPT than in other criteria. Although the minimal gap in the mean difference of vocabulary scores implied that ChatGPT was comparable in identifying the appropriate words and lexical resources, the slight difference in the mean score of the grammar criterion between human raters and ChatGPT indicated a minor leniency of ChatGPT in interpreting grammatical errors compared to human raters. It can be seen that grammar and vocabulary were two aspects in which ChatGPT-graded scores were most closely aligned with human-graded scores. This showed the slight consistency between human raters and ChatGPT in assessing lexical and grammatical elements.

Discussion

The first primary goal of this study was to determine whether ChatGPT's writing assessments differ from those of Vietnamese EFL teachers. The SPSS analysis revealed that although there was no statistically significant difference in grammar and vocabulary scores between ChatGPT and human ratings, significant discrepancies were observed in total scores and in content and organization scores. Therefore, it can be concluded that there was a difference between ChatGPT's writing assessments and those of Vietnamese EFL teachers. This outcome consistently aligns with previous investigations. In particular, Nguyen and Tran (2023), Steiss et al. (2024), Bucol and Sangkawong (2024), Kim et al. (2024), and Bui and Barrot (2025) have clearly demonstrated the contrast between AI-generated and human-based assessments. Interestingly, a smaller group of researchers, such as Uyar and Büyükahıska (2025) and Koraishi (2024), has reported partial alignment between ChatGPT and human evaluations under specific conditions, suggesting that AI assessments were sometimes similar to teacher evaluations, but only in a limited and inconsistent manner. To sum up, the integration of ChatGPT in writing evaluation must be reviewed with caution, as the overall scores confirm clear differences between ChatGPT's writing assessments and those of Vietnamese EFL teachers, despite some alignment in certain contexts.

The second research question aimed to investigate which assessment criteria, including content, organization, grammar, and vocabulary, had the greatest inconsistency between ChatGPT and human writing evaluation. Overall, the findings show a clear misalignment between human raters' scores and ChatGPT's, with ChatGPT tending to give higher scores than human raters across four criteria. This pattern may imply that human raters might apply stricter evaluation standards than ChatGPT when assessing students' writing. A similar tendency was reported by a study of Geçkin et al. (2023). The author noted that human raters are inclined to carefully examine sentence details and the overall quality of the writing. Such differences in evaluative focus between human raters and ChatGPT may partly explain human raters' lower scores. By comparing the mean scores of human raters and ChatGPT for each criterion, the results reveal that ChatGPT was more generous, assigning a 0.5 higher score in the content criterion than humans. It can be concluded that the content criterion shows the greatest discrepancy between

ChatGPT and human ratings, indicating a considerable difference in how human raters and ChatGPT prioritize and interpret content elements. While teachers often evaluate the depth, clarity, and relevance of ideas, ChatGPT may place greater emphasis on the general completeness of the responses. However, this gap should not necessarily be viewed as a limitation in ChatGPT's evaluation of content-related aspects; rather, it is evidence of a distinct focus on writing evaluation between ChatGPT and human raters. Besides that, the organization criterion is the second-largest discrepancy, since ChatGPT assigned a 0.45 higher score than a human. One possible explanation is that ChatGPT-based evaluation may be less sensitive to subtle errors related to the logical flow and coherence of ideas, suggesting that ChatGPT and human raters also differ in their focus on organizational aspects. These findings align more closely with those of previous studies (Jackaria et al., 2024; Topuz et al., 2025; Yoon et al., 2023; Yang, 2024), which also reported a misalignment between ChatGPT and human raters in assessing organizational and content elements. In contrast, ChatGPT demonstrates higher consistency with human raters when evaluating grammar and vocabulary criteria, with scores 0.25 and 0.13 higher than those of human raters, respectively, suggesting a closer level of agreement between ChatGPT and human raters when evaluating linguistic aspects. One possible reason is that grammar and vocabulary are two linguistic features that can be identified more objectively through rules and patterns. While this finding shows that grammatical and lexical features are the two aspects most closely aligned with human-given scores, it contrasts with Poláková et al. (2024), who found that ChatGPT has limitations, especially in accurately assessing grammatical features. These different results indicate that ChatGPT's performance in assessing grammar criteria may vary across contexts, and its reliability in this area remains an issue for further investigation.

Conclusion

The study identified several key differences in essay assessment between ChatGPT scores and those of Vietnamese EFL raters. The results indicated that noticeable variations existed between AI-generated scores and those given by human raters. Particularly, ChatGPT tended to give higher scores than human raters across all criteria. This pattern demonstrates ChatGPT's leniency in writing evaluation, encouraging an understanding of differences in writing evaluative orientation between ChatGPT and human raters rather than an assumption of inaccuracy. Among the four criteria, content and organization showed the greatest gap. ChatGPT has proven generous in assessing content and organizational elements, such as the relevance of ideas, the logical progression of main points, etc. In contrast, the smallest discrepancies between ChatGPT and Vietnamese EFL raters' scores appeared in grammar and vocabulary criteria, indicating that scoring patterns in these two areas of ChatGPT were closely aligned with human raters. In other words, AI-based scoring may align with human evaluations of lexical-related factors and grammatical range and accuracy.

The study's results show that ChatGPT is a promising tool for EFL writing assessment. For large-scale scoring, ChatGPT may help teachers reduce their workload; however, human oversight remains important for assessing content and organizational criteria. To promote the effectiveness of writing assessment practices in Vietnam, a balanced integration of ChatGPT and teachers may help ensure the accuracy and reliability of writing assessment in the Vietnamese educational setting.

Although this study offers many valuable insights to enhance the accuracy of using ChatGPT in writing assessment, some limitations still remain. Firstly, the study's sample size was small, comprising 20 essays written by 20 students at one level in a university. Therefore, the result of

this study may lack generalizability. Secondly, only one essay genre, the opinion essay, was examined in the study, so this result may not be generalizable to other essay types. Further research can enhance generalizability by using a larger sample and involving students from different levels, and students' levels should be assessed using a test designed according to a standardized scale. Moreover, future studies can include more human raters and sample essays, and each essay should be assessed by multiple raters to calculate inter-rater reliability and increase the reliability of human evaluation. Besides, only ChatGPT version 4.0 was used to grade essays, and future versions may yield different results. Therefore, further studies should incorporate multiple versions of ChatGPT to ensure generalizability, as updated versions may yield different results. The last suggestion is to incorporate teacher interviews or questionnaires to better understand teachers' attitudes, concerns, and suggestions regarding the integration of AI tools into language assessment.

Acknowledgments

We received plenty of enthusiastic help and support that guided and encouraged us to conquer all obstacles and finish this study. First of all, we would love to express our thanks to Assoc. Prof. Pham Vu Phi Ho, PhD, at the Industrial University of Ho Chi Minh City. Based on his guidance, encouragement, and sincere comments from the beginning days of this process, we are now mature enough to recognize our hidden potential and reduce our weaknesses. Moreover, I would like to express my sincere thanks to the students from the Academy of Public Administration and Governance in Ho Chi Minh City (APAG) and to the many beloved teachers who took the time to participate in this research. In short, we sincerely thank all the people who are helping us finish this study.

References

Atasoy, A., & Moslemi Nezhad Arani, S. (2025). ChatGPT: A reliable assistant for the evaluation of students' written texts?. *Education and Information Technologies*. https://doi.org/10.1007/s10639-025-13553-1

Baker, N. L. (2014). "Get it off my stack": Teachers' tools for grading papers. *Assessing Writing, 19*, 36–50. https://doi.org/10.1016/j.asw.2013.11.005

Berry, V., Sheehan, S., & Munro, S. (2019). What does language assessment literacy mean to teachers?. *ELT Journal*, *73*(2), 113–123. https://doi.org/10.1093/elt/ccy055

Bucol, J. L., & Sangkawong, N. (2024). Exploring ChatGPT as a writing assessment tool. *Innovations in Education and Teaching International*, *62* (6), 1–16. https://doi.org/10.1080/14703297.2024.2363901

Bui, N. M., & Barrot, J. S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Educ Inf Technol 30*, 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

Geckin, V., Kızıltaş, E., & Çınar, Ç. (2023). Assessing second-language academic writing: AI vs. Human raters. *Journal of Educational Technology and Online Learning*, *6*(4), 1096–1108. https://doi.org/10.31681/jetol.1336599

González, E. F., Trejo, N. P. & Roux, R. (2017). Assessing EFL university students' writing: a study of score reliability. *Revista Electrónica de Investigacion Educativa, 19*(2), 91–103. https://doi.org/10.24320/redie.2017.19.2.928

Hang, N. T. T. (2021). Vietnamese upper-high school teachers' views, practices, difficulties, and expectations on teaching EFL writing. *Journal on English as a Foreign Language*, *11*(1), 1–20. https://doi.org/10.23971/jefl.v11i1.2228

Hoang, T. T. H., Dang, H. N., Pham, N. B. Q., & Truong, T. K. X. (2025). EFL teachers' perceptions of utilizing ChatGPT in designing lesson plans for IELTS reading skills. *International Journal of AI in Language Education*, *2*(3), 19–39. https://doi.org/10.54855/ijaile.25232

Hong, W. C. H. (2023). The impact of ChatGPT on foreign language teaching and learning: Opportunities in education and research. *Journal of Educational Technology and Innovation*, *3*(1), 37–45. https://doi.org/10.61414/jeti.v5i1.103

International English Language Testing System (IELTS). (2023). *IELTS Writing Band Descriptors*. https://ielts.org/cdn/Guides/ielts-writing-band-descriptors.pdf

Jackaria, P. M., Hajan, B. H., & Mastul, A. R. H. (2024). A comparative analysis of the rating of college students' essays by ChatGPT versus human raters. *International Journal of Learning, Teaching and Educational Research, 23*(2), 478–492. https://doi.org/10.26803/ijlter.23.2.23

Kim, H., Baghestani, Sh., Yin, Sh., Karatay, Y., Kurt, S., Beck, J., & Karatay, L. (2024). ChatGPT for writing evaluation: Examining the accuracy and reliability of AI-generated scores compared to human raters. *Exploring artificial intelligence in applied linguistics* (pp. 73–95). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.06

Koraishi, O. (2024). The intersection of AI and language assessment: A study on the Reliability of ChatGPT in grading IELTS writing task 2. *Language Teaching Research Quarterly*, 43, 22–42. https://doi.org/10.32038/ltrq.2024.43.02

Li, J., Huang, J., Wu, W., & Whipple, P. B. (2024). Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings: A preliminary investigation. *Humanities and Social Sciences Communications, 11*(1), 1–9. https://doi.org/10.1057/s41599-024-03755-2

Ludwig, S., Mayer, C., Hansen, C. L., Eilers, K., & Brandt, S. (2021). Automated essay scoring using transformer models. *Psych, 3*(4), 897–915. https://doi.org/10.3390/psych3040056

Nguyen, T. T. H. (2023). EFL Teachers' Perspectives toward the Use of ChatGPT in Writing Classes: A Case Study at Van Lang University. *International Journal of Language Instruction*, *2*(3), 1–47. https://doi.org/10.54855/ijli.23231

Nguyen, T. H. B., & Tran, T. D. H. (2023). Exploring the Efficacy of ChatGPT in Language Teaching. *AsiaCALL Online Journal*, *14*(2), 156–167. https://doi.org/10.54855/acoj.2314210

Palmer, J., Williams, R. E., & Dreher, H. (2002). Automated essay grading system applied to a first-year university subject - How can we do it better?. *Proceedings of IS2002 Informing Science and IT Education Conference* (pp. 1221–1229). Informing Science Institute. https://doi.org/10.28945/2553

Pham, M. T., & Cao, T. X. T. (2025). The Practice of ChatGPT in English Teaching and Learning in Vietnam: A Systematic Review. *International Journal of TESOL & Education*, *5*(1), 50–70. https://doi.org/10.54855/ijte.25513

Poláková, P., Ivenz, P., & Klímová, B. (2024). Examining the reliability of ChatGPT as an assessment tool compared to human evaluators. *Procedia Computer Science, 246*, 2332–2341. https://doi.org/10.1016/j.procs.2024.09.543

Topuz, A. C., Yıldız, M., Taşlıbeyaz, E., Polat, H., & Kurşun, E. (2025). Is generative AI ready to replace human raters in scoring EFL writing? Comparison of human and automated essay evaluation. *Educational Technology & Society*, *28*(3), 36–50. https://doi.org/10.30191/ETS.202507_28(3).SP04

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of Applied Learning and Teaching, 6*(1), 342–363. https://doi.org/10.37074/jalt.2023.6.1.9

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

Uyar, A. C., & Büyükahıska, D. (2025). Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT. *International Journal of Assessment Tools in Education, 12*(1), 20–32. https://doi.org/10.21449/ijate.1517994

Vo, L., & Huynh, N. (2025). Vietnamese EFL Teachers' Perspectives on ChatGPT: A Conceptual Metaphor Analysis. *Arab World English Journal*, *16*(1), 162–178. https://doi.org/10.24093/awej/vol16no1.10

Yang, Y. (2024). The Reliability of Using ChatGPT in Rating EFL Writings. *Shanlax International Journal of Education, 12*(4), 49–59. https://doi.org/10.34293/education.v12i4.7855

Yoon, S. Y., Miszoglad, E., & Pierce, L. R. (2023). Evaluation of ChatGPT Feedback on ELL Writers' Coherence and Cohesion. *arXiv.* https://doi.org/10.48550/arXiv.2310.06505

Biodata

Ho Nhut Nam completed a Bachelor's program in English Language at the Industrial University of Ho Chi Minh City, Vietnam, and is pursuing a master's degree at the same institution. He has many years of teaching young learners and adults. His research interests include the integration of AI into language education, second language acquisition, and interdisciplinary studies connecting language and culture.

Tra Thi Cam Thu is currently pursuing a master's degree in English Language at the Industrial University of Ho Chi Minh City. She is an experienced teacher of English who has worked with a wide range of students at a public school and VUS. Her practical classroom experience has inspired her academic focus on discovering effective strategies that enhance both student engagement and the language learning process for EFL learners..

Pham Thi Quyen is currently pursuing her master's degree in English language at the Industrial University of Ho Chi Minh City, Vietnam. She is an officer at the Department of International Cooperation at a university in Ho Chi Minh City and has three years of experience in teaching English to young learners. Her interest in doing research is in English teaching and learning methodology.

Tran Phuong Thanh is currently pursuing an MA in English Language at the Industrial

University of Ho Chi Minh City, Vietnam, building on her Bachelor of Arts in the same field. She has over three years of experience in teaching English communication courses to young learners and adults at private English centers. Her current academic and professional focus lies in advancing teaching methodology and exploring the effective integration of AI in educational contexts.