# Evaluating ChatGPT's Reliability in Grading Writing Assignments on the EOP Learning Platform

Tran Yen Van[1*] , Le Thi Huong Giang[1]

[1] Hanoi University of Industry, Vietnam
[*]Corresponding author's email: vanty@haui.edu.vn
[*] https://orcid.org/0000-0002-8783-9276
https://doi.org/10.54855/979-8-9870112-8-7_1



Received: 24/06/2025 Revision: 10/09/2025 Accepted: 20/09/2025 Online: 03/10/2025

## ABSTRACT

As Artificial Intelligence (AI) is being used more and more in education, utilizing AI to grade the writing of students is a concern for trustworthiness. Using a mixed-methods research design that combines both quantitative and qualitative data collection tools - questionnaires and semi-structured interviews - this study investigates the reliability of using ChatGPT to mark students' writing assignments on the EOP online learning platform (https://eop.edu.vn/) compared with human evaluators at the School of Languages and Tourism, Hanoi University of Industry. The findings provide the advantages and limitations of AI-supported grading, highlighting the accuracy, consistency, and alignment with human grading criteria of AI grading. The attitudes of teachers toward AI scoring are also examined in this paper to determine its accuracy. Recommendations for enhancing AI scoring systems to enable more effective and fairer assessments are provided based on the findings. The research contributes to the academic literature on the use of AI in education, emphasizing the importance of sustaining the enhancement of AI-driven evaluation tools to enable effective and fairer online learning.

**Keywords**: AI grading, assessment reliability, ChatGPT, writing evaluation, advantages and limitations

## Introduction

The integration of artificial intelligence (AI) technology has increased dramatically in various sectors. Montenegro-Rueda et al. (2023) highlight the significant impacts of AI implementation in education, including improvements such as enhanced operational efficiency, facilitation of global learning, development of smarter learning resources, and optimization of learning management systems. As an example of AI technology, ChatGPT is easily recognizable because it utilizes an advanced language model to convert input data into text that mimics human communication. Originally designed for natural language processing tasks, ChatGPT has found its application in educational contexts, particularly in foreign language learning. However, one should be cautious because the implementation of AI tools may pose certain threats, such as the

CITATION | Tran, Y. V., & Le T. H. G. (2025). Evaluating ChatGPT's Reliability in Grading Writing Assignments on the EOP Learning Platform. *ICTE Conference Proceedings, 7*, 1-19. ISSN: 2834-0000, ISBN: 979-8-9870112-8-7. DOI: https://doi.org/10.54855/979-8-9870112-8-7_1

possibility of unauthorized access to private data and the dissemination of false information (Dwivedi et al., 2023; Floridi & Chiriatti, 2020). These risks highlight the importance of critical oversight and ethical considerations when adopting AI in education. Wang et al. (2024) discuss how AI merges Automation and Learning Technologies to solve conventional problems in education through systems that can adapt learning as well as provide personal tutoring. These applications span from providing wide-ranging assessments and tools for management, which offer immediate feedback, to predictive profiling systems that assist teachers in foreseeing learners' strengths and weaknesses before these issues emerge.

Writing is a repetitive process. It is rare for learners to produce a perfect piece on their first try. The process requires drafting, reviewing, revising, and even writers might start from the beginning (Flower, L., & Hayes, J. R. (1981)). Traditional methods of getting feedback, such as waiting for tutor comments or peer reviews, can be time-consuming (Hyland, K., & Hyland, F. (2006)). As a result, AI's instantaneous response proves effective. Learners can submit a draft and receive feedback immediately, which enables them to address the problems without any delay (Wang et al. (2024)). This ease not only makes the learner's writing process more efficient but also makes it more adaptive and collaborative. The feedback provided by AI tools like ChatGPT is another area where their innovative potential becomes evident. It does not limit itself to grammar checks or vocabulary suggestions, but also handles the sophisticated aspects of academic composition (Montenegro-Rueda et al., 2023). Feedback can be provided in various aspects, such as the depth of content, or it can identify areas where arguments lack clarity or the evidence is still weak. Structural inconsistencies can be highlighted, ensuring that the narrative flow of the essay or paper remains unbroken. The writing style of learners can also be evaluated to ensure it is consistent and tone-appropriate for the target group. However, the most significant aspect of this feedback is its potential to elevate the quality of a learner's thinking in terms of originality and depth (Montenegro-Rueda et al., 2023).

The EOP online learning platform is a self-contained system designed to integrate knowledge learning with the acquisition of the four basic language skills: listening, speaking, reading, and writing. Of these, writing is most important because students are required to submit written assignments upon completing each lesson. This daily writing task enables students to practice new language skills in real-world contexts and develop professional and academic communication skills. The design of the platform facilitates iterative learning by providing space for recursive submission, reflection, and rewriting of texts, reflecting the circularity of writing. This way, ChatGPT can be utilized to provide students with immediate, constructive feedback in all areas of language, with an emphasis on writing skills.

Although AI-based evaluation has several advantages, such as instant, standardized feedback, convenience in handling bulk submissions, and learner-tailored recommendations, it also creates a host of issues. To its credit, AI software is able to identify grammatical mistakes, construction faults, and even offer remarks on coherence of subject matter, thus making the revision process a lot easier and prompting learners to learn in small steps (Wang et al. (2024)). However, concerns regarding the objectivity and validity of AI decisions, especially in deciding creativity, critical analysis, and subtlety of argument, remain. There is also the potential for students to over-depend on AI feedback at the expense of cultivating independent critical

learning abilities. And, of course, technical constraints, including inability to appreciate context or cultural allusions, result in misinterpretable or misguided feedback. These combined elements of AI incorporation render it essential to critically analyze its role and reliability in evaluating student writing within the dynamic online learning context of the EOP platform.

With both the promise and the risks of incorporating AI technologies, such as ChatGPT, into learning settings, a question arises about the extent to which such tools can be trusted to carry out tasks that have traditionally been the preserve of human teachers. Within the context of the EOP online learning context - in which students submit writing assignments on a weekly basis - the question at hand is whether AI-enhanced grading can prove a viable and sound alternative or supplement to human judgment. Although early indications suggest that ChatGPT can provide efficiency, speed, and consistency of response, further research is needed to determine the tool's capacity for assessing more nuanced aspects of writing, such as logic, tone, originality, and argument quality, particularly in specialized or professional areas of discourse. This research aims to investigate the reliability of grading students' writing tasks using ChatGPT on the EOP online learning platform.

## Literature review

### *AI in education*

The expansion of artificial intelligence (AI) in education has triggered unprecedented changes in teaching, learning, and assessing students. Numerous AI-powered learning tools have emerged, supporting various learning processes, including intelligent tutoring systems, adaptive learning software, language learning software, and computer-delivered test systems (Luckin et al., 2016). These tools are designed to enhance educational outcomes, alleviate teachers' workload, and provide students with immediate, personalized feedback based on their performance.

AI-powered systems are increasingly used to monitor student engagement, personalize learning pathways, and support self-regulated learning. As observed by Holmes et al. (2019), the systems in question typically rely on machine learning and natural language processing (NLP) to enable them to perform advanced operations such as text data analysis, identification of learning patterns, and even generating content in real-time.

However, even though there are many benefits of AI in education, several ethical and practical concerns have been raised. As Selwyn (2019) argues, AI systems can unintentionally reinforce pre-existing biases due to data constraints with which they are trained. Secondly, such systems may not always possess the rich contextual and cultural understanding that their human instructor counterparts enjoy, thus raising concerns about fairness and transparency in decision-making. These concerns are further magnified in high-stakes testing environments where automatic systems inevitably have disproportionate influence on students' grades.

This study aims to investigate the reliability and appropriateness in assessing students' writing by comparing the performance of ChatGPT with that of human evaluators at School of Languages and Tourism, Hanoi University of Industry. Additionally, it examines teachers' perspectives on AI-based evaluation and provides recommendations to enhance AI grading systems, aiming for more efficient and equitable online learning environments.

### *AI in grading and writing assessment*

Among the earliest applications of AI in learning is Automated Essay Scoring (AES), a machine learning and natural language processing-driven technology for scoring essays written by students. AES software, such as e-rater (created by ETS) and IntelliMetric, has also been widely applied in massive-scale standardized testing as supportive tools or to wholly replace human raters (Shermis & Hamner, 2012).

However, standard AES systems have their own shortcomings. They will most likely not accurately assess creative, off-the-beaten-track, or analytical writing and are not typically able to detect plagiarism or factual inaccuracies (Ho, L. T. P., Doan, N. A. H., & Dinh, T. L., 2023). As Weigle (2002) states, holistic writing assessment entails subjective judgment—something that current AI models have yet to fully replicate. Consequently, hybrid models that combine AI scoring and human moderation are being experimented with to find a balance between accuracy and efficiency (Nguyen Thi Thu, H. (2023)).

This study compares the accuracy of ChatGPT in grading students' writing on the online learning platform (https://eop.edu.vn) with that of human graders from the School of Languages and Tourism at Hanoi University of Industry. Through a mixed-methods design that combines quantitative analysis of scores with qualitative interviews of instructors, the study examines whether ChatGPT's grades align with human scoring criteria in terms of consistency, accuracy, and fairness.

### *ChatGPT as a tool for writing assessment*

The advancement of large language models such as ChatGPT has significantly expanded the possibilities for using AI in educational assessment, particularly in evaluating writing. As a tool powered by natural language processing (NLP), ChatGPT possesses several features that align well with the demands of writing assessment, including its fluency, ability to provide constructive feedback, consistency in scoring, and potential for scalability in large educational settings.

One of the primary strengths of ChatGPT is its ability to produce coherent, contextually relevant text, making it a suitable tool for assessing the most important aspects of writing, including vocabulary, sentence structure, grammar, coherence, and organization. It also has the ability to give feedback in a quite human-like manner, which is especially handy in formative assessment contexts.

ChatGPT's second main strength is its ability to reduce the inconsistency and fatigue that are typically the hallmarks of human graders. The model has the capability to grade large volumes of writing at standardized levels, a feature that is very useful in high-stakes testing or large-scale testing (Dwivedi et al., 2023). Secondly, ChatGPT can be fine-tuned or guided through the use of specific scoring rubrics, enabling it to deliver more accurate judgments based on specified criteria.

Even with all the benefits, there are giant caveats. ChatGPT is weak at judging creative expression, evaluating argument quality, or determining whether content is task-suitable to the degree that requires human nuance of judgment and situational discrimination (Sari, 2024). ChatGPT may therefore be an effective assistant to facilitate grading but is not yet mature enough to fully replace human graders.

Dwivedi et al. (2023) also discussed the broader applications of generative AI in teaching, emphasizing that ChatGPT can be an effective grading tool, particularly in language-intensive courses. Nonetheless, they cautioned that over-reliance on such systems may result in plausible

but inaccurate feedback or evaluations.

Therefore, a major unanswered question explored in this study is whether ChatGPT can offer reliable and valid writing evaluations in real-world EOP contexts, particularly in comparison with teacher grading on the https://eop.edu.vn platform.

### *Reliability and validity in AI scoring*

In writing assessment, reliability refers to consistency in scoring, whether between different human scorers or between humans and computers. Validity, on the other hand, pertains to the degree to which scores assigned correspond to actual students' writing quality (Weigle, 2002). They are the two fundamental terms in any language assessment, particularly where incorporation of artificial intelligence is part of the assessment process.

Later research presents conflicting opinions regarding the dependability of AI-based writing testing tools such as ChatGPT. Kumar and Rose (2023a), for instance, suggest that ChatGPT can provide relatively consistent ratings for successive submissions, which indicates a positive level of dependability.

Validity issues are especially critical in relation to the extent to which the AI systems comprehend the writing task, genre conventions, and socio-cultural nuances. Human raters, in contrast, draw on pedagogical knowledge, cultural frameworks, and contextual understanding. AI evaluation has the potential to overlook or misinterpret the significant yet nuanced aspects of student writing. This tension lowers the legitimacy of AI-evaluation scores, particularly when measuring upper-level writing skills such as coherence, argumentation, or rhetorical strategy.

Within the framework of this research on online learning environments in this EOP, these issues are thoroughly examined to determine the extent to which grading by ChatGPT aligns with grading by teachers. There needs to be congruence between AI grading standards and teacher grading standards to see whether ChatGPT is feasible as a backup grading alternative for student writing. By testing for reliability (consistency) and validity (correctness and appropriateness), this research aims to present a balanced perspective on the application of AI in assessing academic writing.

### *Attitudes of teachers toward AI scoring*

The attitudes and perceptions of teachers towards AI-based marking systems are instrumental in the adoption and utilization of such technology in schools. Despite most teachers being aware of the potential for AI to provide instant feedback, improve consistency, and reduce workload, concerns persist regarding the accuracy, transparency, and pedagogical responsiveness of AI-powered tests (Ranalli, Link, & Chukharev-Hudilainen, 2017).

In the context of student writing evaluation, numerous studies have demonstrated that teachers are confident in the effectiveness of formative assessment and early grading using AI tools, such as ChatGPT. They are, however, uncertain about AI-only grading without human intervention, particularly in grading more subjective aspects such as tone, originality, coherence, and cultural awareness (Zhai, 2022).

These findings are consistent with the results of the present study. Interview responses of teachers in the School of Languages and Tourism from Hanoi University of Industry revealed a cautiously positive attitude towards using ChatGPT as a complementary assessment tool. Although some viewed it optimistically as a means of assisting teaching, they emphasized the importance of combining AI with human judgment to ensure fairness and contextual insight. Apart from this, teachers also indicated the need for more guidelines and clarity on AI feedback for students' work.

These findings highlight the need for ongoing dialogue among AI manufacturers, educators, and education policymakers to ensure that AI-generated assessment materials are pedagogically suitable, ethically accurate, and responsive to the diverse needs of students and educators.

### *Research questions*

The purpose of this study is to examine the accuracy, consistency, and conformity of ChatGPT to human grading criteria in evaluating students' writing assignments on the EOP online learning platform (https://eop.edu.vn/). To address the research questions, both qualitative and quantitative approaches were used. Data were collected through structured questionnaires and semi-structured interviews with teachers at the School of Languages and Tourism, Hanoi University of Industry. This research study aims to address the following research questions:

- To what extent is ChatGPT accurate and consistent in grading students' writing assignments in terms of message content, communicative achievement, organization, and language across different writing tasks?
- How does ChatGPT's grading align with human assessment, and what are the perceived benefits and drawbacks of AI-based evaluation?

Through addressing these questions, the study has a broader goal of determining whether tools such as ChatGPT and other AI tools can be used safely and effectively as part of assessment practices within online language learning environments. The hypothesis for this research study is that ChatGPT's assessment capabilities will demonstrate both strengths and limitations compared to human evaluators.

## Methods

### *Pedagogical setting & participants*

The study was conducted at Hanoi University of Industry during the first semester of the 2024-2025 academic year. It focused on English teachers for Mechanical Engineering students who are currently working at the School of Languages and Tourism. These teachers, who hold a master's degree and have at least 5 years of teaching experience, are proficient in grading students' writing, particularly for English for Occupational Purposes (EOP). Eighteen teachers participated in the study and completed the questionnaire during the data collection phase. Five of them were then chosen to participate in a subsequent semi-structured interview to gather more information about their individual opinions and expertise regarding the use of ChatGPT as a grading assistant. Each interview lasted approximately ten to fifteen minutes and was recorded for content analysis purposes.

### *Design of the study*

This study adopted a mixed-methods approach (Creswell & Plano Clark, 2017), including a structured questionnaire and semi-structured interviews, to investigate the research questions. Two data collection instruments were employed: (1) A structured questionnaire to gather quantifiable data on teacher perceptions and (2) A semi-structured interview protocol to explore participants' reasoning and experience in more depth. The investigation focused on three main

areas: the accuracy of scoring between humans and computers, consistency of grading across one writing assignment to the next, and the degree of agreement with human ratings on marking criteria for message content, communicative achievement, organization, and language use. These were the major areas of concentration for the research.

### *Data collection & analysis*

#### *Questionnaire*

The questionnaire was designed by the researchers to efficiently collect responses from a large number of teachers. It consists of two main parts. The first section collects demographic information, such as teaching experience and familiarity with AI tools, including different ChatGPT versions (Free, Plus, or Pro). In the second section, a 5-point Likert scale is used to compare ChatGPT's grading to human scoring. Accuracy was rated from 'very accurate' to 'very inaccurate'; consistency was measured using another 5-point Likert scale from 'very consistent' to 'very inconsistent.' The teachers used different ChatGPT plans (Free, Plus, and Pro) depending on their individual access and preferences. ChatGPT was rated by teachers based on four major criteria: language quality, organization, communicative effectiveness, and message content. Questions on the extent to which trust was placed in AI outputs were also included in the survey. Descriptive statistics, including means and standard deviations, were calculated with Microsoft Excel. We grouped open-ended responses into themes to collect initial qualitative insights.

#### *Interview*

While the questionnaire provided quantitative data, interviews were conducted to add depth and detail to the study results. Five teachers were selected to participate in a semi-structured interview, with each teacher assigned a code to ensure confidentiality during data analysis. . To ensure confidentiality, they were assigned pseudonyms (T#1 to T#5). Each interview lasted approximately ten to fifteen minutes and was recorded for content analysis purposes. Twenty interview questions were prepared based on themes emerging from questionnaire responses to gain a deeper understanding of teachers' motivations for their ratings and perceptions regarding the use of AI tools in marking. These interviews helped reveal the weaknesses of AI grading, especially in assessing complex aspects such as the coherence of ideas, expected formats, and relevance to the assigned task. All the interviews were audio-captured and converted into text for later qualitative analysis.

The interviews were transcribed word for word and reviewed for themes using Braun and Clarke's (2006) method. We developed initial codes from the data, and then refined the themes through repeated review. A second coder reviewed a sample of transcripts to verify consistency in coding, and we discussed any discrepancies to resolve them.

### *Instrument validation and evaluation*

To justify the clarity, dependability, and relevance of the instruments, the questionnaire and interview guide were both pilot-tested with two seasoned EOP instructors who were not part of the main study. Minor corrections to word choice were made based on their suggestions and to meet research goals. Experts validated the content validity of the questionnaire, while pilot

testing confirmed its face validity. We checked the internal consistency of the Likert scale items using Cronbach's alpha, which showed a satisfactory measure of reliability (α = 0.81), indicating good internal consistency. For qualitative data, we enhanced credibility through triangulation of sources (interviews and questionnaires), member checking (participants reviewed their interview transcripts), and peer debriefing during coding. These methods helped ensure that the results accurately reflected participants' perceptions and minimized bias.

## Findings and discussion

### *4.1. Accuracy of AI grading (message content, organization, language)*

#### *4.1.1. Accuracy in message content*

Table 1

Assessing writing content: task requirements, idea development, communicative purpose

| Items | **Mean** | **SD** |
|---|---|---|
| AI's assessment accuracy of students' writing task completion | 4.00 | 0.58 |
| AI's assessment of logical coherence in writing | 3.78 | 0.42 |
| AI's assessment accuracy of writing's communicative purpose (appropriate tone and style) | 3.61 | 0.68 |

Table 1 presents the ratings of participants' perceptions of AI accuracy in content assessment for writing, categorized under three sub-criteria: task requirements, development of ideas, and communicative intent. The Mean (M) scores across these criteria range from 3.61 to 4.00 on a 5-point Likert scale, where 1 indicates 'very inaccurate' and 5 indicates 'very accurate'. According to commonly accepted benchmarks, mean scores from 3.41 to 4.20 are interpreted as high. Therefore, the Mean (M) scores across the three criteria are relatively high, reflecting the overall consensus among participants on the suitability of AI applications in this area.

For task requirement assessment, the highest mean of 4.00 was recorded for AI's capability to understand and judge whether students have addressed all task components, with a moderate SD (0.58) suggesting consistent responses among participants. The same phenomenon can be observed in the development of ideas, where a slightly lower Mean of 3.78 (SD = 0.42) suggests that while AI is generally reliable, a percentage of respondents doubt its ability to accurately evaluate logic and coherence in writing. This selection had been explained more from interview with (T#3 and T#5) "*I think AI does quite well. However, logic and coherence in writing are factors that require subtle perception, which AI may not be able to do as well as humans*." and "*I find AI to be an effective support tool, especially for simple, well-structured writing. However, I expect that in the future, AI will improve its ability to recognize nuances and communicative intent to better support teaching.*" (T#2 and T#4)

Interestingly, the assessment of communicative purpose (tone and style) revealed a range of opinions, with the lowest Mean of 3.61 (SD = 0.68). This selection was elaborated upon by

Teacher #1 and Teacher #4 in their interviews "*I appreciate the ability of AI to identify the requirements of the assignment. I have full confidence that AI can support teachers in this part. However, I still want to have human review for complex writing*." (T#1 and T#4); and "*I find AI to be an effective support tool, especially for simple, well-structured writing. However, I expect that in the future, AI will improve its ability to recognize nuances and communicative intent to better support teaching*" (T#2 and T#5). This suggests that AI systems may still struggle with nuanced elements of communication, such as tone appropriateness and stylistic variation.

In general, our results support Ranalli et al. (2017), who suggest that AI tools can be useful for augmenting content and structure analysis, but continue to require human assistance in evaluating more subtle communicative aspects.

### *Accuracy in organization*

Table 2

Assessing the organization in student writing

| | **Mean** | **SD** |
|---|---|---|
| AI's ability to distinguish formal and informal language in writing | 4.11 | 0.57 |
| AI's assessment of the organization (structure and use of cohesive devices) in writing | 3.78 | 0.79 |
| AI's evaluation of logical paragraphing and sequencing of ideas | 3.67 | 0.67 |

The information in table 2 shows that AI exhibited a generally good performance in assessing the organizational aspects of student writing. More specifically, for the use of formal/informal language, the highest mean score was observed (M=4.11, SD = 0.57), indicating strong confidence among participants in this aspect. Further explanation for this selection emerged from the interviews with T#3 and T#5 "*I found the AI to do quite well at distinguishing between formal and informal language. This is a clear strength that helps students develop an awareness of appropriate writing style.*" However, when it came to measuring structural aspects and cohesive elements, the outcome (M = 3.78; SD = 0.79) reflected moderate confidence that AI could manage complex organizational aspects. Similarly, users' measurement of logical paragraph sequencing had the highest mean (M = 3.67) with greater variability (SD = 0.67), implying some uncertainty among users. Additional insights into this choice were provided by T#2 and T#4 during the interviews *"AI can provide initial support in assessing the organization of ideas within paragraphs, especially in simple-structured passages. However, there is still some doubt about the logic between paragraphs."*

These results are consistent with Kumar & Rose (2023b) who found that while AI is in fairly good agreement with human raters on evaluating coherence and elementary structure, differences emerge in the evaluation of more advanced logical organization. The doubt cast by Neutral responses here corroborates Perelman's (2013) condemnation of computer scoring programs for favoring shallow organization over sophisticated logical thinking.

### *Accuracy in language*

Table 3

Assessing the accuracy of grammar and vocabulary in writing

| | **Mean** | **SD** |
|---|---|---|
| AI's accuracy in identifying grammatical errors and sentence structure issues | 4.44 | 0.60 |
| AI's assessment of vocabulary use and its ability to distinguish between minor and major language errors | 3.72 | 0.80 |

Table 3 findings indicate the rating by participants of AI accuracy in identifying the application of grammar and vocabulary in writing. In grammar, the response was very good, with the highest mean value (M = 4.44, SD = 0.60), suggesting that almost all respondents highly rated AI as very reliable in detecting and correcting grammatical errors. This result is also supported by interview responses, such as the one from teacher T#2, who stated, *"AI gives instant correction for grammatical errors, which makes students aware of their habitual mistakes and learn faster."*

However, for vocabulary, the tests differed more. AI's assessment of vocabulary use and its ability to distinguish between minor and major language errors produced a lower mean score (M = 3.72, SD = 0.80). This result suggests that, although AI is widely used to mark vocabulary-related errors, its ability to fully understand nuances in word choice and usage remains open to average doubt by users. Teacher T#4 also shared this view, citing that *"AI sometimes misses context-specific word choices or suggests alternatives that don't always fit naturally."* This is a case for human intervention in vocabulary assessment.

These findings are in support of the arguments by Weigle (2002), who posited that "anything is likely more probable to be rated better where a rubric was used."

### *Consistency of AI grading in student writing*

### *Consistency in task fulfillment and the relevance of the idea*

Table 4

Consistency in task fulfillment and the relevance of the idea

| | **Mean** | **SD** |
|---|---|---|
| AI's consistency in assessing students' fulfillment of task requirements | 3.50 | 0.96 |
| AI's consistency in evaluating the relevance of ideas in student writing | 3.50 | 0.69 |

As can be seen from table 4, both criteria-AI's consistency in assessing students' fulfillment of task requirements and AI's consistency in evaluating the relevance of ideas—received identical mean scores (M = 3.50), indicating a moderate level of confidence among participants. However, the standard deviation was higher for task fulfillment (SD = 0.96), suggesting greater

variation in responses, while the relevance of ideas showed lower variability (SD = 0.69), reflecting slightly more consistent perceptions in that aspect. The primary causes of the disparate perceptions were discovered through follow-up interviews. *"Although ChatGPT works quite well on descriptive tasks, it tends to provide only minimal commentary on argumentative writing"* (T#1), *"It is more reliable when the standards are specific but less reliable when they are general"* (T#4), and *"ultimate decisions still require human judgment, even though it is useful for initial feedback"* (T#5).

These findings align with those of Lu et al. (2024) and Kasneci et al. (2023), which suggest that AI models excel on formal tasks but still struggle with creative or context-rich writing. While ChatGPT can operate reliably in surface-level testing, its shortcomings in deeper analysis suggest that human input remains crucial, especially in open-ended tasks or those involving a specific genre.

### *Consistency in communicative achievement and organization*

Table 5
Consistency in communicative achievement and organization

| | **Mean** | **SD** |
|---|---|---|
| AI's consistency in recognizing appropriate communicative purposes (tone and style) | 3.44 | 0.76 |
| AI's consistency in assessing the organization (structure and use of cohesive devices) in student writing | 3.72 | 0.80 |

Table 5 shows ratings of participants for AI consistency in evaluating communicative organization and achievement in students' writing. AI consistency in evaluating the organization (structure and cohesive device usage) received a higher mean rating (M = 3.72, SD = 0.80) compared to AI consistency in ascertaining communicative purposes (M = 3.44, SD = 0.76), with struggle to adapt to more intricate evaluations. These issues were brought up in teacher interviews: *"ChatGPT was good at straightforward tone and organization but had a problem with complex organization or register shift"* (T#3, T#2). "Final judgment still needs to be human judgment" "clear rubrics helped make the judgment more precise" (T#4, T#5). These findings are in agreement with previous research (Lu et al., 2024; Kasneci et al., 2023), where AI software was proven to be sufficient for the evaluation of writing on the surface but not able to handle more advanced or context-dependent characteristics.

### *Consistency in language (grammar and vocabulary)*

Table 6
Consistency in language (grammar and vocabulary)

| | **Mean** | **SD** |
|---|---|---|
| AI's consistency in detecting grammatical errors in student writing | 4.06 | 0.97 |
| AI's consistency in assessing vocabulary use and language appropriateness | 3.56 | 0.76 |

Data in Table 6 indicate that the consistency of AI in marking grammatical errors attained a fairly high mean (M = 4.06, SD = 0.97), suggesting that participants viewed AI as generally consistent in marking grammatical errors. For the consistency of AI in marking vocabulary usage and language appropriateness, the mean was relatively low (M = 3.56, SD = 0.76),

indicating continued concerns about how AI handles context-sensitive vocabulary usage and subtly nuanced language appropriateness. Interviews validated this impression: educators referred to ChatGPT as *"reliable for elementary grammar but less so for nuanced flaws"* (T#2) or *"contextually appropriate vocabulary in business communication"* (T#3). *"Clear rubrics improved its performance*" (T#4), *"although choices were still better made by humans"* (T#5). These findings are consistent with those of Guo & Wang (2024) and Prompiengchai et al. (2025), who argue that although AI tools can assist in evaluating writing at the surface level, human expertise remains necessary for a more thorough evaluation in EOP contexts.

### *Alignment with human grading criteria in student writing*

### *Task fulfillment and idea development*

Table 7
Assessing the alignment between AI and human scoring in terms of content and idea development

| | **Mean** | **SD** |
|---|---|---|
| AI's alignment with human grading in assessing students' fulfillment of task requirements | 3.61 | 0.49 |
| AI's alignment with human grading in evaluating logical development in writing | 3.67 | 0.74 |

Table 7 presents a comparison of alignment between human and AI scoring for task completion and logical development of ideas. The moderate mean scores were M = 3.61 (SD = 0.49) for task completion and M = 3.67 (SD = 0.74), slightly higher for logical development, reflecting a wide overall but cautious evaluation of the alignment of AI grading. This selection had been explained more from interview with T#3 and T#5 "*Personally, I also feel that AI can handle the basic requirements of ideas and content, but the subtleties of idea development still need human judgment.*"; and "*AI can help with simple, straight-to-the-point writing. But if the writing is complex and analytical, I believe AI is not yet capable of replacing teachers*" (T#2 and T#4). This indicates that while AI scoring is generally perceived as competent in identifying content and idea development, it is not yet fully comparable to human judgment.

These findings align with other studies, which suggest a need for a greater understanding of the partial yet imperfect alignment between automatic systems and raters. This is also supported by Kumar and Rose (2023), who discovered that although ChatGPT exhibits proficiency in surface-level writing assessment, it occasionally misses fine-grained idea flow that human evaluators identify.

Ultimately, it can be said that the findings from the research are in line with prior literature, which suggests that despite AI demonstrating median reliability in content evaluation and idea generation, it is not yet a flawless independent tool for assessment. Human raters are still much needed, particularly when ideas are being assessed for originality and coherence—a finding that seems to be supported by both empirical research studies, such as Ranalli et al. (2022) and Kumar & Rose (2023b), and theory-based critiques, such as those from Perelman (2013) and Selwyn (2019).

### *Communicative Achievement and Organization*

Table 8
Assessing the appropriateness of communicative purpose, style, and organizational structure

| | Mean | SD |
|---|---|---|
| AI's alignment with human grading in assessing the achievement of communicative purpose (appropriate tone and style) | 3.61 | 0.49 |
| AI's alignment with human grading in distinguishing between formal and informal language use | 3.83 | 0.46 |
| AI's alignment with human grading in assessing the overall organization (structure and use of cohesive devices) in writing | 4.00 | 0.58 |
| AI's alignment with human grading in evaluating logical paragraphing and sequencing of ideas | 3.89 | 0.51 |

Table 8 illustrates that the majority of participants rated AI as effective in measuring communicative purpose, style, and organizational structure in writing. Of these, the highest mean was found for overall organization (M = 4.00, SD = 0.58), indicating a high level of agreement that AI efficiently measures structural writing aspects. Logical paragraphing and order of ideas also posted a fairly high mean (M = 3.89, SD = 0.51), indicating positive attitudes. Formal/informal language distinction (M = 3.83, SD = 0.46) and communicative purpose (M = 3.61, SD = 0.49) yielded slightly lower means, indicating that participants exhibited more hesitation in these subtle aspects. Further explanation for this selection emerged from the interviews with T#2 and T#4 *"AI supports well in terms of organization and layout. However, with higher requirements such as emotions and intonation suitable for the reader, I am not completely confident."*

This is similar to Kumar and Rose (2023b), who conclude that AI marking is comparable to human marking for broad features like organization, but deviates significantly when communicative purpose and subtlety of cohesion are involved.

Furthermore, Luckin et al. (2016) argue that AI systems are designed to support human judgment, rather than replace it, especially in matters of communicative appropriateness and creativity. Along the same lines, Perelman (2013) criticized what had transpired in earlier automated scoring systems when the mechanicalness of operations overshadowed meaning.

Generally, the findings substantiate previous studies that AI provides acceptable assistance in evaluating communicative structure but is not complete without human input to conduct an overall analysis of writing.

### *Language (Grammar and Vocabulary)*

Table 9
Assessing the appropriateness of grammar and vocabulary

| | Mean | SD |
|---|---|---|
| AI's alignment with human grading in identifying grammatical errors and sentence structure issues | 4.22 | 0.63 |
| AI's alignment with human grading in evaluating vocabulary use and distinguishing between minor and major language errors | 3.83 | 0.54 |

Table 9 shows participants' evaluations of using grammar and vocabulary. The greatest mean

score (M = 4.22, SD = 0.63) was found for identifying grammatical errors and sentence structure issues, indicating strong agreement that AI performs well in this area. The vocabulary usage and identification of lesser and greater language mistakes had a lesser mean (M = 3.83, SD = 0.54), but with more variation, and indicating a generally favorable view.

This finding aligns with Zhai (2022), who reported that Chinese college students had a positive experience with ChatGPT as a writing tool, particularly for enhancing grammar and vocabulary. The students found the tool beneficial for clarity and grammatical accuracy, although they also recognized the need for post-editing to ensure naturalness and appropriateness in academic writing.

### *Teachers' perspectives on using ChatGPT for assessing students' paragraphs*

The teachers interviewed generally expressed positive attitudes toward using ChatGPT for assessing students' writing, especially paragraph-level tasks. Their perspectives aligned closely with the quantitative data presented in Tables 1 to 9.

### *Advantages*

The analysis of survey data revealed that the most well-known advantage of AI-powered testing is its efficiency in terms of time and speed. The precision of AI to identify grammatical errors and sentence structure flaws was among the strengths the teachers noticed, as indicated by a high mean score (M = 4.22, SD = 0.63). Teacher T#2 affirmed that *"ChatGPT is extremely good at correcting basic grammar errors, which saves teachers' time at first drafts."*

This skill enables students to identify common language problems and assist in self-editing.

Moreover, the capacity to recognize formal and informal language (M = 3.83, SD = 0.46) was highlighted as being particularly useful. Teacher T#5 noted, *"It makes students more aware of tone and register in formal assignments such as reports or proposals."* The trustworthiness of AI organizational feedback (M = 4.00, SD = 0.58) also helped the students develop more rational and coherent writing. Teachers were interested in whether the AI facilitated instant formative feedback without requiring constant monitoring, thereby allowing students to create a series of drafts at a faster pace.

### *Disadvantages*

Despite recognizing the usefulness of ChatGPT in writing assessment, teachers also noted some limitations which can be grouped into three main areas:

- Lexical and pragmatic limitation: Teachers expressed concern over AI's limited ability to evaluate nuanced vocabulary and tone. This was reflected in relatively lower mean scores on areas like judging vocabulary (M = 3.83, SD = 0.54) and judging communicative purpose (M = 3.61, SD = 0.49). As Teacher T#3 put it, "*AI still cannot judge more subtle vocabulary differences or whether the tone completely aligns with the communicative purpose of the text.*"

- Content and contextual understanding: While consistency scores were moderate (M = 3.50 - 3.72) for task fulfillment and idea generation categories, instructors still warned that AI tends to skip context-specific meaning or misinterpret creative ideas at times, especially in more complex or customized writing.

- Pedagogical concerns and student dependency: A further concern that was revealed by T#4 was over-reliance on AI response by students, and it diminishes their critical thinking ability while proof-reading their own work: "*Students are prone to taking AI suggestions blindly without questioning them, which can limit their independent language development*."

Overall, teachers identified ChatGPT as a valuable tool for supporting writing assessment, particularly in terms of grammar, organization, and formality. They noted, however, that AI feedback should complement, not replace, human judgment in assessing content relevance, tone, and the use of sophisticated vocabulary.

### *Implications*

The findings of this study have several important implications for students and instructors when using AI resources, such as ChatGPT, in writing evaluations.

### *For instructors:*

- The fairly high mean scores on grammar identification (M = 4.22, SD = 0.63) and organizational assessment (M = 4.00, SD = 0.58) indicate that ChatGPT can prove to be an efficient support tool in formative assessment activities. Teachers can utilize AI feedback to reduce their marking load, especially for first drafts, allowing them to spend more time on higher-order aspects such as the originality of thought, appropriateness of content, and advanced use of language.

- However, the modest vocabulary and communicative purpose scores (M = 3.61–3.83) suggest that teachers must carefully scrutinize AI feedback in these areas and supply explanatory supplement or correction where necessary. Educating instructors in the interpretation and revision of AI-created feedback will prevent misunderstandings and enhance its value in the classroom.

*For students:* The study highlights how AI provides constructive, immediate feedback on paragraph organization, formal tone, and grammar, enabling students to become more independent during the revision process. Students, however, need to be able to acquire skills in critically evaluating AI feedback without relying on simple recommendations. Teachers need to incorporate activities that allow students to compare AI feedback with human feedback, thereby facilitating critical thinking and independent editing abilities.

### *A recommended hybrid AI-assisted model*

AI-assisted evaluation is considered effective for highlighting surface-level features in writing skills. It serves as a valuable tool for both students and teachers. These systems do great jobs in checking grammar and punctuation, improving sentence structure, and organizing paragraphs. With the help of AI, students can detect the tone and style to know if it is formal or informal, and then ensure it matches the targeted audience. The system can also work in conjunction with detection tools to identify potential plagiarism. The list of services that utilize such functions includes ChatGPT, Grammarly, Turnitin, and Quillbot, all of which offer AI capabilities. In practice, AI-assisted evaluation helps provide comments and feedback on first drafts, offers real-time writing assistance, prepares for peer review, and enables self-editing before final submission. Instructors also assess the strength of argument structure, the clarity of ideas, and

the overall effectiveness of the writing. These activities can serve as steps for students to refine their writing efficiently and effectively.

Meanwhile, evaluations from human instructors are better at solving higher-level tasks that AI tools cannot fully address. Instructors or teachers can have valuable looks into the writing quality and effectiveness. Teachers can focus more attention on specific teaching goals and also evaluate and offer tailored responses on phrasing, word choice, and style, as well as refine areas such as vocabulary. Additionally, they also focus on the clarity of the presented ideas and the integration of various elements of the argument in respect to its depth and balance. Along with evaluating, teachers actively help in understanding the feedback generated by AI. They highlight the limitations of AI and offer practical guidance on utilizing generated text to improve personal work. With this supervision, writing skill development extends beyond basic revision, enabling the refinement of precision, persuasiveness, and intellectual depth.

With the complementary strengths of instructor and AI tools, an integrated model with AI support can be developed, as outlined in the table below, to provide both shallow and deeper-level feedback during various stages of the writing process.

Table 10
A hybrid AI-assisted model

| No. | Steps | Activity | Tool / Actor | Purpose |
|---|---|---|---|---|
| 1 | 1st drafting | Students write the first draft | Students | Create initial outline and the very first draft |
| 2 | Feedback of AI | Students submit to AI for initial feedback | ChatGPT / Grammarly | Identify initial issues |
| 3 | 2nd drafting | Students revise using AI output | Students | Develop learners' autonomy |
| 4 | Classmate and teacher feedback | Classmates and teacher review revised drafts | Teacher+ Classmates | Give feedback deeply |
| 5 | 3rd writing | Students revise using combined insights | Student | Create improvement |
| 6 | Assessment | Teacher grades using rubrics | Instructor | Balance assessment |

## Conclusion

The findings of this study indicate that AI-based assessment demonstrates strong potential in supporting writing evaluation, particularly in areas related to task requirements, idea development, and organizational structure. Survey results showed that a majority of participants rated AI's performance as "Accurate" or "Consistent" in most aspects, reflecting positive perceptions of its efficiency and consistency. Teachers acknowledged AI's role in providing quick feedback, which can effectively aid learners in revising and improving their writing at early stages. However, the study also highlighted limitations in AI's ability to fully assess context-dependent and subjective features of writing, such as tone, communicative appropriateness, and creativity. The presence of a substantial number of "Neutral" responses across these areas, along with interview insights, confirmed that while AI can complement human assessment, it cannot completely replace the role of teachers in writing evaluation.

## Acknowledgement

During the process of preparing and completing this article, we were fortunate to receive valuable guidance, encouragement, and support from many individuals. Their help was essential to the completion of this work.

Additionally, we would like to extend our sincere thanks to our English faculty colleagues for their contributions, encouragement, and helpful feedback. Their support was essential to the completion of this article.

## References

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

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365–387. https://doi.org/10.2307/356600

Ho, L. T. P., Doan, N. A. H., & Dinh, T. L. (2023). An Investigation into The Online Assessment and The Autonomy of Non-English Majored Students in Vinh Long Province. ICTE Conference Proceedings, 3, 41–51. https://doi.org/10.54855/ictep.2334

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Hyland, K., & Hyland, F. (2006). Interpersonal aspects of response: Constructing and interpreting. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 206–224). Cambridge University Press.

Dwivedi, Y. K., Hughes, D. L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and machines*, *30*(4). 681-694 https://doi.org/10.1007/s11023-020-09548-1

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*(7), 8435–8463. https://doi.org/10.1007/s10639-023-12306-1

Kumar, R., & Rose, C. (2023a). The promise and peril of ChatGPT for language assessment. *Language Testing, 40*(2), 123–139. https://doi.org/10.1177/02655322231156807

Kumar, S., & Rose, C. (2023b). Evaluating ChatGPT as a writing evaluator: A comparison with human raters. *Journal of Educational Technology Development and Exchange, 16*(2), 20–35. https://doi.org/10.18785/jetde.1602.02

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Montenegro-Rueda, M., Fernández-Cerero, J., Fernández-Batanero, J. M., & López-Meneses, E. (2023). Impact of the implementation of ChatGPT in education: A systematic review. *Computers, 12*(8), 153. https://doi.org/10.3390/computers12080153

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

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson Education.

Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? *Assessment & Evaluation in Higher Education, 49*(5), 616–633. https://doi.org/10.1080/02602938.2023.2290436

Plano Clark, V. L. (2017). Mixed methods research. *The Journal of Positive Psychology*, *12*(3), 305-306. https://doi.org/10.1080/17439760.2016.1262619

Prompiengchai, S., Narreddy, C., & Joordens, S. (2025). A practical guide for supporting formative assessment and feedback using generative AI. *arXiv*. https://arxiv.org/abs/2505.23405

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of L2 writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology, 37(1), 8-25*

Sari, A. N. (2024). *Exploring the potential of using AI language models in democratising global language test preparation. International Journal of TESOL & Education*, 4(4), 111–126. https://doi.org/10.54855/ijte.24447

Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education*. Polity Press.

Shermis, M. D., & Hamner, B. (2012, April). Contrasting state-of-the-art automated scoring of essays: Analysis. Paper presented at the *Annual Meeting of the National Council on Measurement in Education (NCME)*, Vancouver, Canada.

Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications, 252*, 124167. https://doi.org/10.1016/j.eswa.2023.124167

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Zhai, X. (2022). ChatGPT user experience: implications for education: A review and research agenda. *Educational Technology Research and Development, 70*, 1–24. http://dx.doi.org/10.2139/ssrn.4312418

## Biodata

**Ms. Tran Yen Van** is a lecturer of English at Hanoi University of Industry, Vietnam. She has been teaching for 18 years. During those times, she has been interested in ELT, especially developing students' proficiency in Listening, Reading, Writing and Speaking skills as well as communication skills. Her research interests include Computer Assisted Language Learning, Cognitive Linguistics, Educational Technology, and ELT Methodology.

**Ms. Le Thi Huong Giang** has been teaching English in Hanoi University of Industry since 2007. She received BA degree in 2005 and MA degree in 2010 from the University of Language and International Studies, Vietnam National University, Hanoi. Her areas of professional interest are designing curriculum, course books in a blended learning environment, developing test specification and writing test items. Currently, she administers a team to design teaching and learning materials for students in Faculty of Mechanical Engineering.