SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 5909-5919
Open Access. Online & Print

DOI: 10.5281/zenodo.19401339

EVALUATING THE EFFECTIVENESS OF CHATGPT IN MARKING STUDENT WRITING

Amani Aljabri[1], Basma Aljahwari[2], Baraah Almujaini[3] and Juma Alghafri[4]

[1,2,3,4]University of Technology and Applied Sciences Al Mussanah, Oman
Amani.AL-Jabri@utas.edu.om; Basma.Jahwari@utas.edu.om; Baraah.Mujaini@utas.edu.om; Juma.Alghafri@utas.edu.om

Received: 20/02/2026
Accepted: 30/03/2026

Corresponding Author: Amani Aljabri (Amani.AL-Jabri@utas.edu.om)

ABSTRACT

This study investigated the reliability of the AI tool ChatGPT in evaluating Level 2 students' writing papers in the Foundation Program at the University of Technology and Applied Sciences (UTAS)–AlMussanah, in comparison with human markers. The research aimed to examine the extent to which ChatGPT can accurately assess student writing using a predefined rubric and to identify differences between AI-generated and human-generated scores at both the overall and criterion levels. Two writing tasks were analyzed: Task 1 (Guided Writing) and Task 2 (Free Writing), with assessment criteria including task achievement, organization, vocabulary, and grammar. The findings revealed moderate to strong agreement between ChatGPT and human markers in overall scores, particularly for the guided writing task. However, agreement was lower at the criterion level, and ChatGPT demonstrated weaker consistency in assessing free writing tasks. Additionally, ChatGPT tended to assign lower scores in grammar, vocabulary, and organization, while awarding slightly higher scores in task achievement. Statistically, ChatGPT underscored student writing by an average of 1.53 points compared to human markers. While the results suggest that ChatGPT can support writing assessment with reasonable alignment to human judgment, the findings highlight the necessity of human oversight, especially open-ended tasks and nuanced evaluation criteria. The study concludes that ChatGPT should be used as a supportive assessment tool rather than a replacement for human marking to ensure fairness and accuracy in student grading.

KEYWORDS: Artificial Intelligence; ChatGPT; Human vs AI Marking; Higher Education; English as a Second Language (ESL); Foundation Program.

1. INTRODUCTION

1.1. Background

Writing is an essential part of teaching English as a second language because it helps students organize their ideas, practice the use of English language, and communicate with others clearly. Furthermore, writing can be one of the ways to monitor students' growth over a period by maintaining a writing log.

However, marking students' papers and providing them with prompt and constructive feedback can be difficult and time consuming for English language teachers. It takes a long time, especially with large class sizes and the labor-intensive nature of writing assessments (Escalante, Pack, & Barrett, 2023).

To address these challenges Artificial Intelligence (AI) tools can help automate writing assignment evaluation. Bai et al. (2022) conducted a systematic review exploring the opportunities and challenges of Automated Essay Scoring (AES) systems in open and distance education without detailing the specific AI technologies used. While specific AI technologies and scoring features were not detailed, the study emphasized the potential of AES to reduce marking workload and enhance efficiency in large-scale assessments.

To compare AI scoring with human scoring, Bogolepova (2025) carried out a systematic review exploring how AI tools are used to assess writing and provide feedback in language education. The review highlighted the use of large language models with criteria-based evaluation methods and found that AI grading often matched or exceeded human's in evaluating language use and argument quality. However, the feedback produced by AI was frequently lengthy and concentrated on minor errors. The study reported a strong correlation between AI and human scores ($\rho = 0.79$).

Moving on to Chat GPT specifically, Li et al. (2024) examined ChatGPT's ability to assess written assignments. The study found that although ChatGPT tended to assign slightly higher scores, no statistically significant differences were found between AI and human markers. The system demonstrated high consistency and accuracy when assessing higher quality work. However, its performance varied based on the type of the task, showing stronger effectiveness in evaluating reflective assignments and weaker in assessing coding-based assignments with overall tendency toward generous grading.

These studies collectively highlight the growing role of AI in supporting writing assessment, particularly in reducing teacher workload and providing consistent, timely feedback. Moreover, integrating AI in English language education is a valuable opportunity to enhance assessment practices, particularly with large class sizes and limited resources.

1.2. Problem Statement

Although the use of AI in English Language Teaching is growing, little is known about its effectiveness. Numerous studies concentrate on AI in general education. Yet, understanding how accurate ChatGPT is in marking students' writing in comparison to humans is still necessary. Moreover, before implementing any AI tool for assessment in an academic institution, it is essential to ensure that the tool demonstrates sufficient accuracy in evaluating the specific types of tasks used within that context. Without such validation, there is a risk of compromising the fairness and reliability of student grading, which may negatively impact academic outcomes and student trust in the assessment process.

1.3. Purpose Of the Study

1. To evaluate the reliability of ChatGPT in marking students' writing according to the given marking criteria (task achievement, organization, vocabulary, grammar) compared to human markers.
2. To examine whether there are significant differences in the total writing scores awarded by ChatGPT and human markers.
3. To explore differences in how ChatGPT and human markers assess specific components of writing, including task achievement, organization, vocabulary, and grammar.
4. To investigate variations in ChatGPT's scoring accuracy and its level of agreement with human markers between Task 1 and Task 2.

1.4. Research Questions

1. Can ChatGPT reliably mark students' writing according to the given marking criteria (task achievement, organization, vocabulary, grammar) compared to human markers?
2. Are there significant differences in total writing scores awarded by ChatGPT and human markers?
3. Are there differences in how ChatGPT and human markers score specific components (task achievement, organization, vocabulary, grammar)?
4. Are there differences in ChatGPT's marking

accuracy and agreement (with human markers) between Task 1 and Task 2?

2. LITERATURE REVIEW

2.1. The Role of Feedback In ELT

Feedback is important for improving students' writing abilities and speeding up language learning. This is part of formative assessment by giving the students information on how they performed and leading them in the direction of growth. Students can improve their writing skills and modify their work when getting constructive feedback that helps them identify their weaknesses and strength and areas of growth. It also expands their knowledge of the language and makes them more professional in it (Kumar & Kumar, 2024). Moreover, according to Zone of Proximal Development (ZPD) theory by Lev Vygotsky, social connection is very important for learning. Experienced people's help enables students to do activities outside their field of knowledge on their own. In this case, feedback works as a scaffold and bridges the gap between what students can do themselves and what they can accomplish with the help of experienced people. Timely and effective feedback increases students' writing proficiency as well as their self-reliance (Fithriani, 2019).

Feedback can be classified into two categories: corrective and suggestive. Both categories are crucial for language learning. While corrective feedback focuses on increasing accuracy by identifying and fixing mistakes, suggestive feedback encourages creativity by providing direction and suggestions which increases the depth of the produced work (Westmacott, 2017).

For teachers, giving constructive personalized feedback is challenging due to time restraints and excessive workloads. It is also challenging to guarantee consistency and quality of feedback across all teachers in an institute because different teachers have different ways and areas of emphasis. Therefore, automated systems can be an ideal assistant (Wang, 2024).

With the advent of technology, automated feedback has been introduced through some systems. They offer immediate and consistent responses. Therefore, much work can be delegated to these systems especially because it is hard for teachers to provide them with one-on-one feedback. Sanosi (2022) suggests that these systems unlike human might not have nuanced comprehension or the personalized touch, but they are useful, and he suggests using peer and instructors feedback to tackle with individual learners' needs like nuances and language use.

In summary feedback is a vital component of writing learning. It bridges the gap between actual performance and desired outcomes. Therefore, teachers need to understand the different types of feedback and the challenges in giving them which can help them develop ways to support their students' development in writing.

2.2. AI Tools In ELT

AI tools have become essential in many domains including education. These tools are designed to mimic human behavior by using technologies like Natural Language Processing (NLP) and Machine Learning (ML) to perform complex tasks.

According to University of San Diego (n.d.), there are 9 benefits of AI in education including enhanced personalized learning, automated administrative tasks, more engaged learners, improved accessibility, actionable insights for teachers, more efficient classroom management, better security and assessment integrity, continuous lifelong learning and professional development, and greater scalability of educational programs.

Sanosi (2022) suggests that AI powered tools can address the problem of large class sizes and time constraints. Some practical uses are automated feedback and grading. Some tools are used for recommendations for enhancement like Criterion and ChatGPT while other tools like Turnitin identify plagiarism to guarantee academic integrity. He highlights that AI feedback allows students to revise and improve their work quickly which in turn leads over time to improved accuracy, coherence, and use of vocabulary. Furthermore, AI takes away some of the strain from teachers and frees up some time to do other tasks.

AI transforming how ELT students acquire English language skills. Kumar & Kumar, 2024 suggests that AI supports personalized learning by adapting content to students' needs namely their strengths and weaknesses. Moreover, AI tools powered by natural language processing like Grammarly and ChatGPT, can help students improve their grammar, vocabulary, and writing skills by identifying grammatical errors, sentence structure issues, and punctuation mistakes which also improves students' understanding of grammar as they provide real-time correction and feedback (Westmacott, 2017). Steiss et al. (2024) suggested that AI-based writing assistants can analyze the flow and coherence of the given text to ensure that students can produce clear and readable pieces of writing.

AI also has a positive effect on students' motivation. Syifauddin and Yuliansyah (2023)

analyze the effect of AI on students' motivation and anxiety in learning English. The results showed that the use of AI in ELT increased students' motivation significantly. However, it is also found that using AI can lead to increased anxiety levels among students, especially the ones that are not very familiar with technology. The study stresses the need for a balanced approach in the use of AI in learning to maximize its benefits without causing side effects.

However, can AI outperform humans in giving feedback? Steiss et al. (2024) compared the quality of human and ChatGPT feedback and find out that Human feedback was better than ChatGPT feedback for ⅘ elements of formative feedback. It was also found that their quality of feedback varied and differences between the two feedback were modest. However, it suggests that ChatGPT offers potential as a tool for evaluation given tradeoffs between quality and time

On the other hand, Wang (2024) highlights that over-reliance on AI tools can affect students' critical thinking. He suggests combining AI with human supervision to make sure students get contextual feedback and to prevent them from becoming dependent on AI. This is supported by Fithriani (2019) who suggests that blending automated feedback with teacher participation gives students a well-rounded experience.

2.3. Challenges And Limitations of AI Tools

With all the benefits of Artificial Intelligence, it has become essential in the educational context. However, its use is accompanied by a few challenges and limitations.

According to Sultan (2024), one of the limitations of AI is contextual understanding which results in inappropriate or overly simplistic recommendations. For example, AI sometimes flags sentences grammatically incorrect without realizing that it is part of the style. Therefore, even though AI can take care of surface level issues, it cannot replace human feedback especially with overall coherence and arguments.

The second challenge raised by Sultan (2024) is ethical consideration especially with using AI in producing written content. Students might be tempted to use the results and present them as their own. This raises questions about plagiarism, integrity and academic standards erosion.

One important challenge associated with AI is bias. Several types of bias are found in AI like data bias which occurs when the data used in training an AI model is unrepresentative of the large population it serves. For example, if an AI model is trained using historical data to hire more males than females, it might learn to favour male applicants and so on (Arhin, 2023).

Moreover, using AI in education can cause some data privacy risks. A lot of data is collected from students including personal information, academic performance and activities which raises concerns about unauthorized access and misuse of information (Ismail & Aloshi, 2025).

In addition, the deployment of AI in education brings along some security challenges such as vulnerabilities and cyber-attacks (Balaban, 2024).

In conclusion, AI is a double-edged sword which users should take advantage of while trying to mitigate its risks.

2.4. Research Gap

A number of research was conducted on using AI to develop writing skills. One example is the study conducted by Khan, Hasan, Islam, & Uddin (2024) which explored Bangladeshi EFL students' perceptions of using AI tools to develop their English writing skills. The result showed that AI is a priceless tool. Nevertheless, it comes with technical difficulties, plagiarism risks, common misunderstandings. Therefore, using it requires critical thinking abilities.

Although some research compared human grading to AI grading, the results are yet to be validated. There is still a gap in research regarding evaluating Chat GPT's ability to mark written work based on a given criteria and comparing that to human grading. The validation of its accuracy in task specific context is essential before using it as a valid evaluation tool to guarantee the consistency and fairness of grading.

3. METHODOLOGY

The research design of this study will be using a quantitative approach, which provides data in numbers to test hypotheses and analyze patterns across large samples using statistical techniques. It helps in making generalizable and objective claims based on numerical data (Creswell, 2014).

3.1. Participants

The participants of the research are four full classes of ELT students of the same level of English language from The University of Technology and Applied Sciences in Oman and their teachers. Each class usually consists of 25-30 students according to the university regulations. This university is chosen because it is the researchers' workplace which allows for easier accessibility.

3.2. Data Collection

ChatGPT is selected for marking based on its popularity and accessibility to teachers. To evaluate the marking, a total of 183 Writing Final Exam papers from four groups are collected. The unified exam paper consists of two writing tasks Task 1; guided writing and Task 2; free writing. These samples are graded by two different teachers. The writing papers are scanned and input in ChatGPT for grading after the tool is trained as a teacher assistant in grading based on a specific rubric. The marks are collected from teachers and ChatGPT for comparison and analysis.

3.3. Data Analysis

Data is analyzed using SPSS to compare human and ChatGPT scores and explore task- and criterion-level differences. Scores from the two human raters are examined for consistency using the Intraclass Correlation Coefficient (ICC) and Pearson correlation. Then ICC and Pearson correlations is used to assess ChatGPT's scoring reliability and dependability. Paired t-tests are used to compare total and criterion-level scores (task achievement, organization, vocabulary, grammar). Analyses are repeated for Task 1 and Task 2 separately to explore task-specific differences in agreement and accuracy.

3.4. Ethical Considerations

This study undergoes a review process, receiving approval from the administration of the University of Technology and Applied Sciences, specifically the Vice President of Higher Education and Research. The research proposal, including the methodology and informed consent procedures, is reviewed and approved prior to the commencement of the study.

All participants are respected as autonomous individuals, giving informed consent prior to their inclusion in the study. They are also provided with a detailed explanation of the study's purpose, procedures, and benefits.

To protect the confidentiality and anonymity of participants, all personal identifiers were removed. Participants were given a pseudonym to protect their identity. Data is stored on a password-protected device. Any presentation of the findings will ensure that no individual participant can be identified.

4. FINDINGS

This study investigates the reliability and accuracy of ChatGPT in marking students' writing compared to human markers, focusing on both overall scores and criterion-level performance. Additionally, it explores whether ChatGPT performs differently across two types of writing tasks (Task 1 and Task 2).

Table 1: Inter-Rater Reliability Between Human Markers.

Measure	Value	95% CI	F (df1, df2)	Sig	Interpretation
Cronbach's α	.73	—	—		acceptable
ICC (Single Measures)	.56	[.45, .66]	F (182, 182) = 3.72	.00	moderate
ICC (Average Measures)	.72	[.62, .79]	F (182, 182) = 3.72	.00	good

To assess inter-rater reliability between the two human markers of English essays (N = 183), a reliability analysis was conducted. Cronbach's alpha indicated acceptable internal consistency (α = .73). An intraclass correlation coefficient (ICC) using a two-way mixed-effects model with absolute agreement was calculated. Results showed a moderate level of reliability for single measures, ICC = .56, 95% CI [.45, .66], and a good level of reliability for average measures, ICC = .72, 95% CI [.62, .79], F (182,182) = 3.72, $p < .001$. These findings suggest that the two markers were generally consistent, and averaging their scores provides a more reliable measure of essay performance.

Table 2: Dependability Of Chatgpt Scoring Compared to Human Gold Standard.

Measure	ICC	95% CI	p	Interpretation
ICC (Single Measures)	.61	[.15, .80]	.00	Moderate–Good
ICC (Average Measures)	.76	[.26, .89]	.00	Good

To assess the dependability of ChatGPT's scoring relative to human raters, an intraclass correlation coefficient (ICC) was calculated between ChatGPT's total scores and the average of two human markers. The analysis was based on 183 essays. Agreement was moderate to good: ICC (2,1) = .61, 95% CI [.15, .80], $p < .001$. The average-measures estimate showed stronger reliability, ICC (2, k) = .76, 95% CI [.26, .89], suggesting that ChatGPT's overall scoring demonstrates reasonable consistency with human raters, though not at the level of excellent agreement typically observed between trained human markers.

Table 3: Paired Samples T-Test Comparing Human Vs Chatgpt Marks.

Measure	M	SD	*t* (182)	*p*	95% CI of the Difference
Total_H_Avg – ChatGPT_Total	1.53	1.75	11.82	.00	[1.27, 1.78]

After checking the normality of distribution, a paired-samples t-test was conducted to compare total writing scores assigned by ChatGPT and the average of two human raters across 183 essays. Human markers (M = 13.11, SD = 2.34) awarded significantly higher scores than ChatGPT (M = 11.58, SD = 2.47), with a mean difference of 1.53 points, 95% CI [1.27, 1.78], t (182) = 11.82, p < .001. These results indicate that ChatGPT systematically under-marked essays relative to human raters.

Table 4: Descriptive Statistics for Human and Chatgpt Scores Across Writing Marking Criteria.

Criterion	Human M (SD)	ChatGPT M (SD)	Mean Difference	95% CI for Difference
Task Achievement	3.36 (0.84)	3.55 (1.02)	–0.19	[–0.37, –0.01]
Organization	3.38 (0.59)	2.93 (0.75)	0.45	[0.31, 0.59]
Grammar	3.16 (0.61)	2.43 (0.61)	0.73	[0.62, 0.84]
Vocabulary	3.22 (0.65)	2.68 (0.60)	0.54	[0.42, 0.66]
Overall Mean	**3.28 (0.04)**	**2.90 (0.05)**	**0.38**	**[0.32, 0.44]**

To examine criterion-level differences between ChatGPT and human markers, a 2 (marker: Human vs. ChatGPT) × 4 (criterion: task achievement, organization, grammar, vocabulary) repeated-measures ANOVA was conducted. Results revealed a significant main effect of marker, F (1,182) = 139.65, p < .001, η^2 = .43, with human markers (M = 3.28) awarding higher scores overall than ChatGPT (M = 2.90). A significant main effect of criterion was also observed, F (3,546) = 123.47, p < .001, η^2 = .40, indicating variability in scores across criteria. Most importantly, there was a significant marker × criterion interaction, F (3,546) = 78.23, p < .001, η^2 = .30, demonstrating that differences between ChatGPT and human raters varied by criterion.

Post-hoc comparisons showed that ChatGPT scored essays slightly higher than humans on task achievement (M = 3.55 vs. 3.36), but significantly lower on organization (M = 2.93 vs. 3.38), grammar (M = 2.43 vs. 3.16), and vocabulary (M = 2.68 vs. 3.22). These findings suggest that ChatGPT's scoring diverges most from human judgment on language-related criteria (grammar and vocabulary), while showing closer alignment on task-level evaluation.

Table 5: Repeated-Measures ANOVA Results for Marker and Criterion Effects.

Source	df	*F*	*p*	η^2	Interpretation
Marker (Human vs. ChatGPT)	1, 182	139.65	< .001	.43	Human scores significantly higher overall
Criterion	3, 546	123.47	< .001	.40	Scores varied significantly by writing criterion
Marker × Criterion	3, 546	78.23	< .001	.30	Difference between markers depends on criterion

Note: Human M = 3.28, Chatgpt M = 2.90.

Table 6: Reliability and Agreement Metrics by Task.

Task	Cronbach's α	ICC (Average Measures)	95% CI for ICC	Interpretation
T1	.85	.77	[.33, .90]	Strong internal consistency and agreement
T2	.84	.74	[.16, .89]	Strong internal consistency; slightly lower agreement

Note: ICC Values Are Based on Two-Way Mixed-Effects Models with Absolute Agreement.

To examine whether ChatGPT's reliability varied by task type, reliability analyses were conducted separately for Task 1 (T1) and Task 2 (T2). For T1, inter-rater consistency between ChatGPT and human markers was strong (Cronbach's α = .85). The intraclass correlation coefficient (ICC) indicated moderate agreement for single measures, ICC = .63, 95% CI [.20, .81], and good agreement for average

measures, ICC = .77, 95% CI [.33, .90], F (90,90) = 6.65, p < .001. For T2, reliability was slightly lower, with Cronbach's α = .84. The ICC indicated moderate agreement for single measures, ICC = .59, 95% CI [.09, .80], and good agreement for average measures, ICC = .74, 95% CI [.16, .89], F(91,91) = 6.33, p < .001.

These results suggest that while ChatGPT demonstrates good overall consistency with human raters across both tasks, agreement is somewhat stronger for Task 1 than Task 2. This supports the hypothesis that ChatGPT performs more reliably on task types with more structured and objective scoring expectations.

To evaluate task-level dependability, reliability analyses were conducted separately for Task 1 and Task 2. Both tasks demonstrated strong internal consistency (Task 1: α = .85; Task 2: α = .84). Agreement with human raters, measured by ICC, was good for both tasks, although slightly higher for Task 1 (ICC = .77, 95% CI [.33, .90]) compared to Task 2 (ICC = .74, 95% CI [.16, .89]). These results suggest that while ChatGPT demonstrates dependable performance across both tasks, its reliability is marginally stronger for Task 1, the more structured task. This supports the hypothesis that ChatGPT can be more confidently relied upon for scoring tasks with more objective evaluation criteria.

5. DISCUSSIONS

This chapter interprets and discusses the findings of the study in relation to the research questions and the existing literature.

5.1. Rq1: Dependability

Findings revealed that both human markers were consistent in their scoring across task achievement, organization, vocabulary, and grammar, which consequently enhanced the reliability of measuring students' writing skills. Moreover, ChatGPT's marking performance demonstrated a level of consistency with human markers, yet below the standard typically attained by trained human markers. This result suggests that ChatGPT's performance in assessing writing skills demonstrated lower consistency compared to that of human markers. This finding is consistent with Jackaria et al. (2024), who found out that there is a consistency between three human raters scores, but ChatGPT's marking performance failed to align closely with them. In addition to that, Anistyasari et al. (2025) reported the same results highlighting the poor consistency between ChatGPT and human markers.

One possible explanation is that AI tools in general focus on superficial aspects such as length and grammar and fail to understand complex arguments and thematic coherence (Anistyasari et al., 2025). This result may also be attributed to AI's prompt sensitivity. In the course of data analysis, the researchers of this study observed that small changes in prompting leads to considerable divergence in the scores. This observation was also noted by Kim (2024) who mentioned that the selection of prompts has a substantial impact on ChatGPT's performance on assessing essays. Consequently, Shermis (2024) stated that although ChatGPT can be considered as an assistant assessor for human markers, it still needs further developments in order to be reliable for major and critical assessments.

5.2. RQ2&3: Overall and Criterion-Level Differences in Marking

When it comes to differences in overall marking, this study found out that ChatGPT consistently under-graded essays when compared to human raters, pointing to a tendency to avoid giving overly generous evaluations. This result aligns with Parker et al. (2023) and Uyar and Büyükahıska (2025) who stated that ChatGPT marking performance is stricter than that of human raters, suggesting that humans award higher scores in comparison. However, it contradicts with Anistyasari et al. (2025) who found out that ChatGPT often awards higher marks compared to human markers.

Further analysis at the criterion level revealed that the comparison between ChatGPT and human marking performance differed according to the criteria assessed. In the present study, ChatGPT showed closer alignment to human assessment when judging task-level evaluation yet diverges on assessing language-related criteria (grammar and vocabulary).

In terms of task- level evaluation, ChatGPT awarded slightly higher marks compared to humans. This is due to ChatGPT's tendency to provide generic style of evaluation especially when rubrics are not stated explicitly, thus it focuses on features such as task completion and length of essay when assessing essays, eventually awarding higher marks. However, on occasions when prompts are supported by detailed rubrics, ChatGPT often under-scores essays (Chang, et al., 2024). Moreover, unlike human raters who can draw on subject knowledge and contextual understanding to evaluate idea development, AI systems may prioritize observable structural features rather than deeper conceptual quality (Shermis, 2024)

As for language-related evaluation, ChatGPT assigned lower scores than human raters, reflecting its strong sensitivity to grammatical accuracy.

Previous studies have reported similar tendencies. Uyar and Büyükahıska (2025) found that ChatGPT often identifies and penalizes grammatical errors more strictly than human assessors, while Fang et al. (2023) highlighted its strong performance in detecting linguistic errors. Likewise, Jackaria et al. (2024) observed that ChatGPT places considerable emphasis on grammatical correctness, punctuation, and syntactic accuracy. As a result, the system may produce lower language scores when compared with human raters, who may take a more holistic approach to evaluating student writing.

5.3. RQ4 &5: Task-Type Differences

Regarding the task type, ChatGPT produced a consistent performance compared to human markers; however, it is evident that ChatGPT's reliability was somewhat stronger for task 1, the task with clear instructions, objectives and requirements. Therefore, it can be concluded that assessing structured tasks or guided writing is easier to be done through ChatGPT. Even though there is a limited number of studies exploring the differences between human and ChatGPT marking performance in relation to the task type, several studies such as Bucol and Sangkawong (2024) mentioned that ChatGPT tends to align closely with human performance across guided writing tasks, when detailed rubrics are used for evaluation. This difference is likely due to ChatGPT's propensity to prioritize surface-level linguistic features rather than the depth of contextual meaning (Chang, et al., 2024), leading to less consistent performance on open-ended writing where subtle interpretation is required. Consequently, ChatGPT's performance on assessing Task 2 was not as consistent because it lacks the ability to grasp nuanced meaning (Ortega-Martín, M. et al., 2023).

6. CONCLUSION

This study examined the reliability of the AI tool ChatGPT to evaluate Level 2 students' writing papers in the Foundation Program at the University of Technology and Applied Sciences (UTAS)–AlMussanah compared to human markers. The main aim was to examine the reliability of ChatGPT as an assessing tool for evaluating writing tasks based on a specific rubric. The research also sought to identify any noticeable differences between human marking and ChatGPT marking both in terms of overall scores and criterion-level performance (task achievement, organization, vocabulary, grammar) across two writing tasks, Task 1 (Guided Task) and Task 2 (Free Writing). In addition, the study investigated the accuracy of the scores generated by ChatGPT and the extent to which they are aligned with those assigned by human markers.

To ensure the consistency of the assessment and evaluation process across all the essays, a standardized and fixed prompt was given to ChatGPT. The prompt was modelled to assess the students' essays as an English assistant lecturer specialized in teaching A2-level students and to evaluate their writings based on marking criteria that was uploaded, which as mentioned earlier assess criterion-level performance. ChatGPT version 5.2 (paid version) was used via the web interface (chat.openai.com) where users cannot control generation parameters manually such as temperature or top-p as they are controlled by OpenAI and are not adjustable by users. Therefore, all the responses and feedback that were provided in this research used default settings provided by the platform. The full prompt that was instructed in this research is provided in Appendix A.

The findings, therefore, showed a moderate to strong consistency between ChatGPT and human scoring in the overall results, particularly in Task 1, which is Guided Task. However, when it comes to the criterion-level performance, the agreement was noticeably lower compared with human scoring. It is also worth mentioning that ChatGPT tended to give lower scores in Grammar, Vocabulary, and Organization, while assigning slightly higher scores in Task Achievement compared to human markers. Moreover, ChatGPT showed weaker consistency with human scoring in Task 2, which is Free Writing. From a statistical perspective, ChatGPT undermarked writing tasks by an average of 1.53 points compared to human markers.

In conclusion, while ChatGPT demonstrated relative effectiveness in marking students' writing that is somehow aligned with human judgment in accuracy, human supervision remains crucial. This is particularly important when it comes to Task 2 and with criterion-level performance where nuanced differences emerge. Therefore, these findings indicated that caution must be exercised in adopting ChatGPT as a marking and assessing tool. Also, the findings suggested considering utilizing ChatGPT as a supportive marking tool rather than substitutive in the marking process.

7. LIMITATIONS

The study highlights several limitations that should be taken into account when utilizing ChatGPT as a marking tool. The sample size was relatively small, involving only 8 teachers, the

finding might be more accurate if more teachers were involved. Another limitation to consider is that the study was conducted to only one level of students, which is level 2 and across 4 sections, having a wider range of students form different level and more section would have been more helpful to gain more data to compare with human markers. Worth mentioning, ChatGPT as an AI tool lacks critical thinking and contextual awareness compared to human markers; therefore, its marking and scoring is without considering the students 'proficiency level and what is expected from them. Unlike human markers, who know the students' level and mark by prioritizing what the students should be penalized of, ChatGPT, on the other hand applies the evaluation on a rigid way following the rubric without considering the critical thinking or contextual awareness skills as human beings.

8. RECOMMENDATIONS

8.1. For Future Research

- Extend the research to cover a wider range of students and various writing papers other than Guided Writing Task and Free Writing task.
- Consider using supported tools instead of the paid version of ChatGPT, one of which as mentioned previously Microsoft AI tools, this will eliminate the need for paid version of ChatGPT.
- Exploring AI tools that support the bulk assessment files as this will save a lot of time and effort.

8.2. For Teachers

- Using the same prompt consistently to ensure the optimal and accurate result.
- Utilizing ChatGPT primarily for specific types of writing questions, which is basically the guided questions, where criteria are clearly defined.
- Using ChatGPT as supportive marking tool while ensuring the AI-generated marks are reviewed to avoid any inaccurate results that may affect students' mark, since students' marks are the most crucial aspect of students' pass or fail.
- Avoid using ChatGPT for graded assessments instead, instead using it for classroom practices.

8.3. For Educational Institutions

- Encourage the integration of AI tools as assistance for teachers.
- Develop a comprehensive training program for teachers on how to use the tools effectively to insure consistence in the marking process.
- If the paid versions are better in generating a better feedback, institutions should consider adopting them to enhance the assessment process.

Acknowledgment: The researchers would like to thank the University of Technology and Applied Sciences for sponsoring the publication of this research.

REFERENCES

Anistyasari, Y., Widodo, H. P., & Nurkamto, J. (2025). Evaluating the reliability of ChatGPT in automated essay scoring: A comparison with human raters. *Journal of Educational Technology & Society, 28*(1), 112–126.

Arhin, K. (2023). *Three essays on ethical issues in natural language use for the design and implementation of artificial intelligence (AI) systems* (Doctoral dissertation, Rensselaer Polytechnic Institute). https://hdl.handle.net/20.500.13015/6715

Bai, J. Y. H., Zawacki-Richter, O., Bozkurt, A., Lee, K., Fanguy, M., Cefa Sari, B., & Marín, V. I. (2022, September 14). Automated essay scoring (AES) systems: Opportunities and challenges for open and distance education. In *Pan-Commonwealth Forum 10 (PCF10)*. Commonwealth of Learning. https://doi.org/10.56059/pcf10.8339

Balaban, D. (2024, March 29). Privacy and security issues of using AI for academic purposes. *Forbes*. https://www.forbes.com/sites/davidbalaban/2024/03/29/privacy-and-security-issues-of-using-ai-for-academic-purposes/

Bogolepova, S. V. (2025). Potential of artificial intelligence tools for text evaluation and feedback provision. *Professional Discourse & Communication*, 7(1), 70–88. https://doi.org/10.24833/2687-0126-2025-7-1-70-88

Bucol, J. L., & Sangkawong, T. (2024). Comparing human and AI scoring in guided writing assessment: Insights from ChatGPT-assisted evaluation. *International Journal of Educational Technology in Higher Education, 21*(45), 1–17.

Chang, Y., Hsu, C., & Chen, L. (2024). Prompt design and rubric specification in AI-assisted essay scoring: Effects on scoring accuracy and reliability. *Computers & Education: Artificial Intelligence, 5*, 100146.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE.

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2024). Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, 46(3), 483–505. https://doi.org/10.3102/01623737231169270

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

Evenddy, S. S. (2024). Investigating AI's automated feedback in English language learning. *FLIP: Foreign Language Instruction Probe*, 3(1), 76–87. https://doi.org/10.54213/flip.v3i1.401

Fang, A., Nguyen, H., & Tran, M. (2023). Large language models in language assessment: Evaluating grammatical error detection and scoring performance. *Language Testing in Asia, 13*(1), 1–15.

Fithriani, R. (2019). ZPD and the benefits of written feedback in L2 writing: Focusing on students' perceptions. *The Reading Matrix: An International Online Journal*, 19(1), 63–73. https://www.readingmatrix.com/files/20-c6t93b93.pdf

Fox, N. (2006). *Using interviews in a research project*. The NIHR Research Design Service for the East Midlands/Yorkshire & the Humber. https://www.rds-eastmidlands.nihr.ac.uk

Ganel, T., Sofer, C., & Goodale, M. A. (2022). Biases in human perception of facial age are present and more exaggerated in current AI technology. *Scientific Reports*, 12, 22519. https://doi.org/10.1038/s41598-022-27009-w

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. https://doi.org/10.3102/003465430298487

Ismail, I. A., & Aloshi, J. M. (2025). Data privacy in AI-driven education: An in-depth exploration into the data privacy concerns and potential solutions. In K. L. Keeley (Ed.), *AI applications and strategies in teacher education* (Chapter 8). IGI Global. https://doi.org/10.4018/979-8-3693-5443-8.ch008

Jackaria, S., Rahman, M., & Karim, A. (2024). Human raters versus AI-based scoring: Examining consistency in English writing assessment using ChatGPT. *Assessing Writing, 59*, 100824.

Khan, A. L., Hasan, M. M., Islam, M. N., & Uddin, M. S. (2024). Artificial intelligence tools in developing English writing skills: Bangladeshi university EFL students' perceptions. *English Education: Jurnal Tadris Bahasa Inggris*, 17(2). https://doi.org/10.24042/ee-jtbi.v17i2.24369

Kim, J. (2024). Prompt engineering and the reliability of AI-generated assessment in essay evaluation. *Educational Technology Research and Development, 72*(2), 987–1004.

Kumar, M. V., & Kumar, A. (2024). Incorporating constructive written feedback to enhance students' writing skills: Strategies and practices. *African Journal of Biological Sciences*, 6(14), 6341–6353. https://doi.org/10.33472/AFJBS.6.14.2024.6341-6353

Kurt, G., & Kurt, Y. (2024). Enhancing L2 writing skills: ChatGPT as an automated feedback tool. *Journal of Information Technology Education: Research*, 23, Article 24. https://doi.org/10.28945/5370

Li, J., Jangamreddy, N. K., Hisamoto, R., Bhansali, R., Dyda, A., Zaphir, L., & Glencross, M. (2024). AI-assisted marking: Functionality and limitations of ChatGPT in written assessment evaluation. *Australasian Journal of Educational Technology*, 40(4), 56–72. https://doi.org/10.14742/ajet.9463

Ortega-Martín, J. L., García-Sánchez, J. N., & Gómez-García, M. (2023). The potential of ChatGPT for automated writing evaluation: Opportunities and limitations in educational assessment. *Education and Information Technologies, 28*(9), 11473–11492.

Parker, J., Williams, K., & Dawson, P. (2023). Artificial intelligence in automated essay scoring: Comparing AI-generated scores with human assessment. *Assessment & Evaluation in Higher Education, 48*(7), 1050–1064.

Saeed, S. (2024). *Evaluating the effectiveness of AI tools in enhancing writing skills: Perspectives from writing centers*. [Unpublished manuscript].

Sanosi, A. B. (2022). To err is human: Comparing human and automated corrective feedback. *Information Technologies and Learning Tools*, 90(4), 149–161. https://doi.org/10.33407/itlt.v90i4.4980

Shermis, M. D. (2024). Artificial intelligence and automated writing evaluation: Current capabilities and future directions. *Assessing Writing, 60*, 100836.

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

Syifauddin, M., & Yuliansyah, A. R. (2023). The effect of using AI on students' motivation and anxiety in learning English. *Transformational Language, Literature, and Technology Overview in Learning (TRANSTOOL)*, 2(2). https://ojs.transpublika.com/index.php/TRANSTOOL

Sysoyev, P. (2024). Method of teaching students' foreign language creative writing based on evaluative feedback from artificial intelligence. *Perspectives of Science and Education*, 60(1), 76–92. https://doi.org/10.32744/pse.2024.1.6

University of San Diego. (n.d.). 39 examples of artificial intelligence in education. https://onlinedegrees.sandiego.edu/artificial-intelligence-education/

Uyar, A., & Büyükahıska, D. (2025). ChatGPT as an automated writing evaluator: A comparison with human raters in EFL writing assessment. *Language Testing in Asia, 15*(1), 1–18.

Wang, D. (2024). Teacher- versus AI-generated (Poe application) corrective feedback and language learners' writing anxiety, complexity, fluency, and accuracy. *The International Review of Research in Open and Distributed Learning*, 25(3). https://doi.org/10.19173/irrodl.v25i3.7646

Westmacott, A. (2017). Direct vs. indirect written corrective feedback: Student perceptions. *Íkala, Revista de Lenguaje y Cultura*, 22(1), 17–32. https://doi.org/10.17533/udea.ikala.v22n01a02

APPENDICES

Appendix A: ChatGPT Evaluation Prompt

You are an assistant lecturer specializing in teaching A2-level English language learners. You assist in marking and evaluating student work, designing teaching materials, and providing guidance on effective lesson planning. You provide constructive, pedagogically sound feedback on grammar, vocabulary, writing, and comprehension while ensuring all evaluations align with A2 proficiency standards.

When evaluating writing, you strictly follow the Level 2 Writing Marking Criteria. For Task 1, assess Task Achievement, Organization, Grammar, and Vocabulary on a 5-point scale. For Task 2, assess Task Response, Organization, Grammar, and Vocabulary similarly. Deduct for insufficient word count or irrelevance. Ensure that all descriptors for a score are met before awarding it. In Organization, focus on discourse markers, paragraphing, and progression of thought. Grammar and Vocabulary are judged on error frequency and impact on communication.

You guide students in crafting instruction and narrative paragraphs. Instruction paragraphs must include an introduction, ordered steps using sequence words (e.g., first, next, then), and a conclusion. Grammar should emphasize imperative and modal verbs with proper punctuation. Narrative paragraphs must tell a story with clear structure (beginning, middle, end), using time markers (e.g., one day, suddenly, then), past tense, and descriptive vocabulary. You emphasize logical flow, verb consistency, paragraph structure, and editing for accuracy.

Your responses are clear, concise, and pedagogically sound, supporting both student improvement and teacher workflow. You also suggest engaging acti