International Journal of Learning, Teaching and Educational Research
Vol. 23, No. 2, pp. 478-492, February 2024
https://doi.org/10.26803/ijlter.23.2.23
Received Jan 4, 2024; Revised Feb 23, 2024; Accepted Mar 5, 2024

A Comparative Analysis of the Rating of College Students' Essays by ChatGPT versus Human Raters

Potchong M. Jackaria*, Bonjovi H. Hajan and Al-Rashiff H. Mastul
Mindanao State University – Tawi-Tawi College of Technology and Oceanography
Tawi-Tawi, Philippines

Abstract. The use of generative artificial intelligence (AI) in education has engendered mixed reactions due to its ability to generate human-like responses to questions. For education to benefit from this modern technology, there is a need to determine how such capability can be used to improve teaching and learning. Hence, using a comparative−descriptive research design, this study aimed to perform a comparative analysis between Chat Generative Pre-Trained Transformer (ChatGPT) version 3.5 and human raters in scoring students' essays. Twenty essays were used of college students in a professional education course at the Mindanao State University – Tawi-Tawi College of Technology and Oceanography, a public university in southern Philippines. The essays were rated independently by three human raters using a scoring rubric from Carrol and West (1989) as adapted by Tuyen et al. (2019). For the AI ratings, the essays were encoded and inputted into ChatGPT 3.5 using prompts and the rubric. The responses were then screenshotted and recorded along with the human ratings for statistical analysis. Using the intraclass correlation coefficient (ICC), results show that among the human raters, the consistency was good, indicating the reliability of the rubric, while a moderate consistency was found in the ChatGPT 3.5 ratings. Comparison of the human and ChatGPT 3.5 ratings show poor consistency, implying the that the ratings of human raters and ChatGPT 3.5 were not linearly related. The finding implies that teachers should be cautious when using ChatGPT in rating students' written works, suggesting further that using ChatGPT 3.5, in its current version, still needs human assistance to ensure the accuracy of its generated information. Rating of other types of student works using ChatGPT 3.5 or other generative AI tools may be investigated in future research.

Keywords: ChatGPT; essay writing; generative AI; human raters; inter-rater

[*] Corresponding author: Potchong M. Jackaria, *potchongjackaria@msutawi-tawi.edu.ph*

1. Introduction

The emergence of generative artificial intelligence (AI), such as Chat Generative Pre-Trained Transformer (ChatGPT), has impacted different fields, including education. A review of the literature on the possible effects of ChatGPT in teaching and learning has shown various viewpoints, ranging from favorable to unfavorable (Rahman & Watanobe, 2023). Among the generative AI tools, ChatGPT is the most popular natural language processing (NLP) program, with a wide acceptance rate among users, reaching one million in just seven days from its launching (Buchholz, 2023). With this, ChatGPT has reached an unprecedented feat among any consumer-based applications, leading to its outright banning in schools by some countries and school districts (Sabzalieva & Valentini, 2023). In 2023, a group of 600 scientists, including many world-renowned technology founders such as Elon Musk and Steve Wozniak, called for a slowdown of development of AI until clear rules and policies have been formulated, indicating the huge concern in its use (Morozov, 2023).

In education, the concerns stem from the ability of ChatGPT to generate human-like responses, such as writing essays, operating computer programs, or solving difficult mathematical problems, to name a few. Since using ChatGPT does not require students to think, this may hamper their critical thinking skills (Fuchs, 2023). For education to benefit from the full potential of generative AI tools, caution is indispensable, underscoring that educators should be critical and aware of their limitations and potential biases. Therefore, the integration of AI models for classroom purposes must undergo rigorous security and privacy considerations (Kasneci et al., 2023). There is also the need to re-evaluate its environmental, regulatory, and ethical requirements, suggesting that generative AI must be used in conjunction with ongoing human monitoring, guidance, and critical thinking (Kasneci et al., 2023).

On the other hand, some scholars have argued that generative AI will be part of people's daily lives, hence its integration in schools being essential. In the same way that teachers and students are using calculators in class, ChatGPT will be an indispensable tool for daily writing and work (Harunasari, 2023). Others have contended that educators and learners should take advantage of the available capabilities of AI tools, such as ChatGPT, rather than forego their use completely (Sharples, 2022). Instead of exclusion, educators should be the ones modelling the best practices for students by incorporating AI tools into classwork and the curriculum (Trust et al., 2023).

The use of large NLP models such as ChatGPT in education is a promising development that offers many opportunities to augment the learning experience of students. It may help foster engagement (Sharma & Yadav, 2023) as well as provide assistance and support (Biswas, 2023; Firat, 2023; Gill et al., 2024). For teachers, generative AI may serve as an assistant in creating lessons and activities, including assessment, thereby freeing teachers from the arduous teaching-related task to focus their instructional time on teaching (Kopp & Thomsen, 2023; Mondal et al., 2023).

As a newly developed field, generative AI in education is a part of the ongoing development process (Imran & Almusharraf, 2023). The majority of the published works on AI in education raise concerns as well as recognize its potentials (Adiguzel et al., 2023; Li et al., 2023; Livberber, 2023). Consequently, researching and finding better ways to integrate ChatGPT into classrooms will not only contribute to the body of literature but will also be of practical importance to educators and learners.

Using the essays of third year pre-service teachers from the Mindanao State University – Tawi-Tawi College of Technology and Oceanography, this study was conducted to determine how ChatGPT may be used in grading students' written works and to discover its consistency compared to human raters. This research topic is of particular interest to us, being members of a teacher education institution. By ascertaining the consistency of ChatGPT compared to human raters in grading student works, it is hoped that this study will advance novel insights that will shed light on the seamless adoption and integration of ChatGPT in the context of teaching and learning.
Specifically, the study sought to answer the following questions:

1) How do human raters versus ChatGPT rate student essays?
2) What is the reliability of human and ChatGPT rating of student essays?
3) What is the inter-rater reliability of human and ChatGPT rating of student essays?
4) Is there a significant difference between human and ChatGPT rating of student essays?

2. Literature Review

In recent years, there has been a rise in the development and application of advanced AI technologies, which has significantly influenced many fields, including education. One such technology is ChatGPT, a large language model developed by OpenAI. While ChatGPT offers exciting opportunities for students and educators alike, it also poses many threats to the realm of traditional education and research.

Educators' perceptions toward the use of generative AI such as ChatGPT are still mostly negative. One study found that the majority of the faculty members involved in their survey exhibited negative perceptions and attitudes toward the use of ChatGPT in the classroom. Reasons cited include potential misuse, such as plagiarism and cheating (Iqbal et al., 2022). In addition, a number of limitations in relation to the use of ChatGPT have been identified, including reduced critical thinking, potential for plagiarism, risk of misinformation, lack of originality and innovation, and limited access to literature (Liyberber, 2023). Studies have suggested that there is a need for more information and education about ChatGPT and generative AI among teachers in order to make informed decisions about its use.

While the ongoing discourse on ChatGPT is generally positive, many are apprehensive in terms of its potential misuse in the area of academic integrity, actual impact on learning and skills development, current limitations and capabilities, lack of existing policy, and social concerns (Li et al., 2023). ChatGPT must then be used with caution, since it can generate human-like responses, creating a possibility for students to plagiarize academic works (Bitzenbauer, 2023; Lo, 2023). Students may instruct ChatGPT to write essays, literary pieces, or computer codes and to solve difficult mathematical problems with detailed explanation and submit them as their own. Since students do not need to think with the use of ChatGPT, this may lead to student over-dependence on the technology, inhibiting critical thinking (Fuchs, 2023; Vargas-Murillo et al., 2023).

Another concern on the use of generative AI is in relation to ethical and practical issues. Other important concerns to solve include the possibility of biases in AI algorithms, where the mainstream opinion is most likely favored. There is also a need to provide teachers with needed preparation and support (Adiguzel et al., 2023). In addition to the ethical issue, there are also concerns on the accuracy of NLP models. While the accuracy may generally be quite good, there are still errors that occur in interpreting meanings or creating accurate information (Lund et al., 2023).

Despite the potential risks posed by ChatGPT, many potential benefits for education are likewise notable. AI may improve learning outcomes, student productivity, and engagement in learning tasks (Adiguzel et al., 2023; Iqbal et al., 2022). It can also be used as a medium for personalized feedback and learning tutorials. Likewise, teachers can use ChatGPT to engage with their students, offer personalized feedback, create interactive conversations, prepare lessons and assessments, and understand new ways to teach complex concepts (Iqbal et al., 2022; Rahman & Watanobe, 2023).

In addition, generative AI tools may help mitigate the perennial problems faced by teachers. One such significant challenge is limited time and resources. Teachers have a limited timeframe within which to manage multiple tasks, such as lesson preparation, grading, and classroom management, leading to stress and burnout, which can ultimately affect the quality of teaching (Kopp & Thomsen, 2023). If these tools are proven to be helpful, teachers may now focus on other important aspects of teaching, such as classroom management and personalized learning. Additionally, ChatGPT can assist teachers in providing instant feedback to students. ChatGPT may be used to analyze student responses in typed format and provide immediate feedback, highlighting areas in which students need to improve. Furthermore, generative AI tools such as ChatGPT can help teachers who are facing limited resources by reducing the need for expensive textbooks and other classroom materials, as it can generate personalized content for teaching–learning purposes (Mondal et al., 2023).

In the area of academic research, ChatGPT could serve as a useful tool for academic writing, such as in the drafting of academic articles. ChatGPT has the capabilities to partially aid in literature reviews by generating ideas in response to natural language inputs and can help with organizing contents. It can also be used as a formatting, editing, and proofreading tool for article polishing (Livberber, 2023; Mondal & Mondal, 2023). ChatGPT can be used as a writing assistant by educators and researchers. However, scholars have warned that it is necessary to lay down a clearer understanding of its role as an aid and facilitator for both the learners and instructors (Imran & Almusharraf, 2023). Hence, academic institutions need to revisit and update their policies on students' and teachers' conduct in the use of generative AI tools. Teacher training and assessment practices in writing courses need to be upgraded to include academic integrity issues, such as plagiarism, originality issues, assignments, and online and home-based exams, factors which can be compromised when students use generative AI tools (Imran & Almusharraf, 2023; Waltzer et al., 2023).

The current version of ChatGPT relies on text-based inputs and can only respond through text. Hence, some researchers have attempted to explore how to use it in writing. It was found that there are opportunities and challenges in using ChatGPT as a writing assistant (Imran & Almusharraf, 2023). Similarly, ChatGPT was found to be a good writing tool and that it can be useful in academic writing. However, it should be used with caution, as writers still need to validate the information it generates (Dergaa et al., 2023; Lingard, 2023).

The use of technology such as NLP tools in automating scoring of student work has been investigated in some studies. In these studies, the results were compared with the scoring by human raters for consistency. For instance, Liu et al. (2015) concluded that feedback from NLP tools on students' essays highly related with that of human raters. In a similar study, McNamara et al. (2015) found that the automated model Coh-Metrix has a 92% adjacent accuracy with human scores. The results indicate that this can be a promising tool for use in scoring student essays. On the other hand, in their study, Dikli and Blyle (2014) found that there is a discrepancy between the human feedback and the automated essay scoring, with the humans providing better quality feedback. ChatGPT, as a new form of generative AI, was examined in the study of Parker et al. (2023). The study concluded that ChatGPT demonstrates utility as an automated writing evaluation tool, though it was found to be slightly stricter than human evaluators.

Based on the literature reviewed, an ongoing debate remains as to whether generative AI tools such as ChatGPT pose an opportunity or a threat to education. Despite the discourse at hand, however, the potential role of ChatGPT in education is unequivocal, benefitting both students and teachers in the teaching and learning process. From the above literature, it can be argued that educational institutions need to adopt new policies on the use of AI and look for ways on how the technology can best be integrated into classrooms. Furthermore, while some studies have been conducted on ChatGPT as a writing evaluation tool, it should be noted that these studies were undertaken in different contexts. Therefore, this study was conducted to determine the consistency of ChatGPT in rating student essays as compared to human raters. The findings in this regard might be potentially useful to classroom teachers, administrators, and future researchers.

3. Methodology

3.1 Research Design

This study utilized a descriptive–comparative research design, a non-experimental way of comparing two variables, where the researchers do not manipulate any of the variables but only describe the sample (Siedlecki, 2020). We made use of quantitative data from the ratings of three human raters, comparing them to the ratings generated by ChatGPT 3.5. ChatGPT version 3.5 was chosen since it is OpenAI's publicly available generative AI tool considered the most popular to date. Hence, considering ease of access, ChatGPT 3.5 is more likely to be used by teachers and students compared to ChatGPT version 4.0, which requires subscription.

3.2 Participants

The participants of the study were 20 college students enrolled at the College of Education of Mindanao State University – Tawi-Tawi College of Technology and Oceanography, a public university in the southern Philippines. The participants were in their third year and represented various education degree programs. Essay writing is usually part of the class activities of education students, which is why they were selected. As to the individual participants, they were selected using the stratified sampling technique, with half of the participants' essays selected randomly from above the mean and the other half from below the average in the pre-rated scores. This was done to provide varied inputs for both human raters and ChatGPT, useful for testing of consistency. To ensure the objectivity of the ratings and the anonymity of the participants, their names were coded during the rating of the essays. Prior to their participation, informed consent was sought from the participants.

3.3 Research Instrument

The main instrument used in this study was the writing rubric for essays adapted from Carrol and West (1989), utilized and proven to be useful for rating using the Delphi technique by Tuyen et al. (2019). The scoring rubric has four criteria, graded separately, namely content (relevance of ideas), organization (coherence and structure), language use (vocabulary and grammar), and mechanics use (punctuation and spelling). Each criterion was given an equal weight of four points, with 1 being *poor* and 4 being *excellent to good*. This rubric was useful due to its advantage of being comprehensive and specific enough to capture the writing skills of students in the given task. Because of the minor revision on the rubric, specifically on the assigned points and levels of interpretation, the adapted rubric underwent validation by three experts and was subjected to reliability testing prior to the actual use. A meeting was held among the experts to discuss the results of the reliability testing. From the suggestions, the improvements in the rubric, specifically on giving of equal points for each criterion, were made. The rubric was then used by the human raters to score the participants' essays. The three human raters were faculty members of the college where the data were obtained, having had more than 10 years of teaching experience and being knowledgeable on the topic discussed in the essay.

The same rubric was used as a prompt inputted into ChatGPT. We used OpenAI's ChatGPT version 3.5 to serve as AI rater. In addition, aside from its popularity, this version of ChatGPT is free for use and hence most accessible to teachers and students. ChatGPT's utility and ease of use are among the reasons for its popularity among students (Alrishan, 2023).

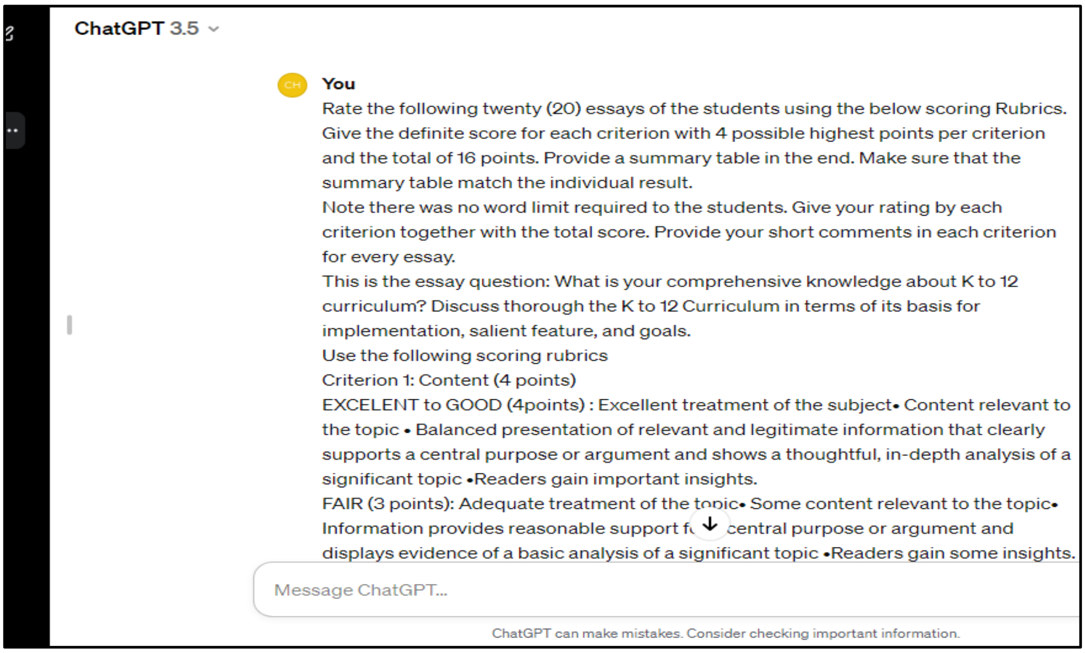
3.4 Data Gathering Procedure

The essays used in this study were taken from the participants' written works in one of the second researcher's classes, a professional education course on curriculum development. The participants were timed and monitored during the writing of the essay to elicit natural writing. Using the pen-and-pencil test, the objective of the essay assignment was for the participants to demonstrate comprehensive knowledge about the K to 12 Curriculum in the Philippines, highlighting its bases for implementation, salient features, and goals. No word limit was set for participants in writing their essays. Because the student essays were written on a sheet of paper, we encoded the essays verbatim as Microsoft Word files before submitting it for rating.

A total of 28 student essays were graded and used in this study. Of these, 8 were used to establish the reliability of the rubric, while the other 20 were used in the final data analysis. To ensure varied results, student essays representing different ability levels were selected. Half of the student essays were above average and the other half below average in the pre-rated scores. While this study used an established rubric, a trial using eight student essays was made to ensure that the three human raters had common interpretations of the rubric. Prior to rating, the human raters were briefly oriented on the context of the study, the essay question, and the criteria of the used rubric. Finally, the essays were rated by the human raters independently.

For the ChatGPT ratings, we started with prompts (see Figure 1). It took us four trials before obtaining the desired response from ChatGPT. In these trials, improvements on the prompts were made, such as inclusion of the essay question, emphasis on the number of essays as ChatGPT had missed some of them, and a request for a summary table. The prompt question and the rubric were inputted to ChatGPT together with the students' essays. Then, the responses were screenshotted and recorded for statistical analysis. This process was done three times with an interval of five minutes between each.

Finally, after the initial quantitative results were available, the three human raters convened to discuss and perform an in-depth analysis of the obtained data. The discussions and inputs were then used to substantiate the prior quantitative findings. This was also to ensure that the findings and its implications were rigorous, drawn from a shared interpretation.



You

Rate the following twenty (20) essays of the students using the below scoring Rubrics. Give the definite score for each criterion with 4 possible highest points per criterion and the total of 16 points. Provide a summary table in the end. Make sure that the summary table match the individual result.

Note there was no word limit required to the students. Give your rating by each criterion together with the total score. Provide your short comments in each criterion for every essay.

This is the essay question: What is your comprehensive knowledge about K to 12 curriculum? Discuss thorough the K to 12 Curriculum in terms of its basis for implementation, salient feature, and goals.

Use the following scoring rubrics

Criterion 1: Content (4 points)

EXCELENT to GOOD (4points) : Excellent treatment of the subject• Content relevant to the topic • Balanced presentation of relevant and legitimate information that clearly supports a central purpose or argument and shows a thoughtful, in-depth analysis of a significant topic •Readers gain important insights.

FAIR (3 points): Adequate treatment of the topic• Some content relevant to the topic• Information provides reasonable support f central purpose or argument and displays evidence of a basic analysis of a significant topic •Readers gain some insights.

ChatGPT can make mistakes. Consider checking important information.

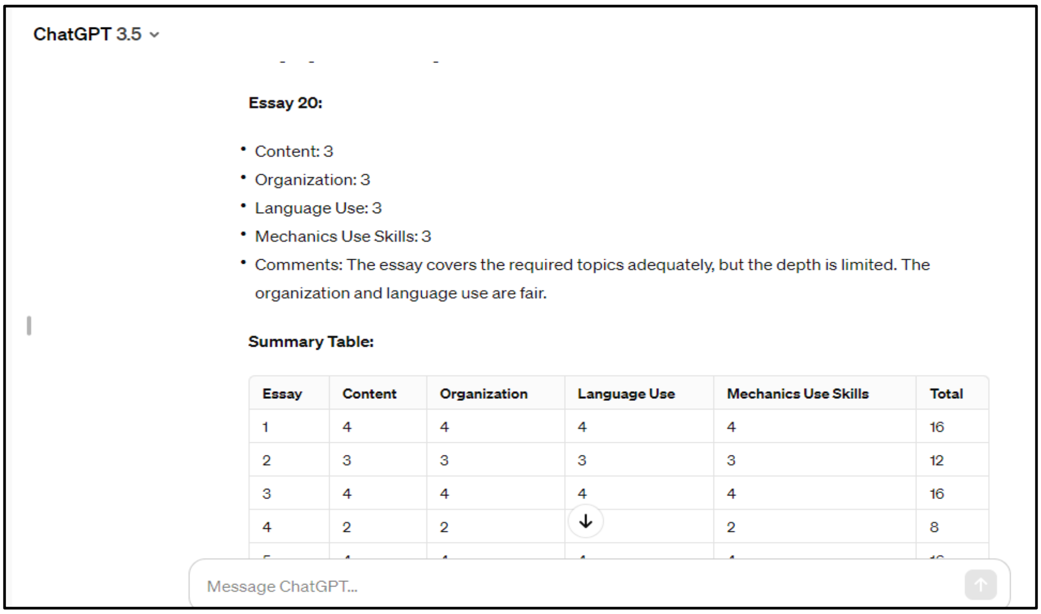
Figure 1: Prompt inputted to ChatGPT 3.5 to rate the student essays

3.5 Data Analysis

Twenty written essays were collected from the participants. Subsequently, we encoded the essays into Microsoft Word file format and then forwarded the file together with the rubric to the three human raters. Simultaneously, we utilized ChatGPT version 3.5 to rate the 20 student essays using the prompts and the rubric as inputs. The written responses of ChatGPT were then copied, taking notes of the ratings per criterion. To find agreement among the three different human raters and the three ratings of ChatGPT 3.5 within the group, the intraclass correlation coefficient (ICC) was utilized. The study used ICC estimates at 95% confidence intervals based on a mean-rating ($k = 3$), absolute-agreement, and two-way mixed-effects model. Similarly, the ICC was used to determine the consistency of the human and ChatGPT ratings by criterion. ICC is a more meaningful indicator in inter-rater agreement (Laschinger, 1992). The interpretation of the results was based on the interpretation scale proposed by Koo and Li (2016) at the 95% confidence interval of the ICC estimate, such as poor (below 0.5), moderate (0.5−0.75), good (0.75–0.9), and excellent (0.90 and above).

However, the ICC only showed the linear consistency of the two data sets compared. Even though the two data sets were different, it was possible to obtain a high reliability coefficient if the difference was consistent. Hence, the Wilcoxon signed-rank test was further utilized. Finally, the results were tabulated, analyzed, and discussed.

Figure 2 provides a sample rating of ChatGPT version 3.5 of student essays based on the rubric.



Essay 20:

- Content: 3
- Organization: 3
- Language Use: 3
- Mechanics Use Skills: 3
- Comments: The essay covers the required topics adequately, but the depth is limited. The organization and language use are fair.

Summary Table:

Essay	Content	Organization	Language Use	Mechanics Use Skills	Total
1	4	4	4	4	16
2	3	3	3	3	12
3	4	4	4	4	16
4	2	2		2	8

Figure 2: Sample rating of ChatGPT version 3.5

4. Results and Discussion

4.1 Human and ChatGPT 3.5 Ratings of Student Essays

Table 1 shows the ratings of the human raters and ChatGPT of the 20 student essays using the rubric from Caroll and West (1986).

Table 1: Means and interpretations of the human and ChatGPT 3.5 essay ratings

Rubric criterion	Human raters			ChatGPT 3.5		
	Mean	SD	Interpretation*	Mean	SD	Interpretation*
Content	2.78	0.64	Fair	2.93	0.39	Fair
Organization	2.48	0.48	Average	2.68	0.50	Fair
Language	2.62	0.41	Fair	2.72	0.41	Fair
Mechanics	2.53	0.38	Fair	2.95	0.70	Fair
Average	**2.60**	**0.48**	**Fair**	**2.82**	**0.50**	**Fair**

**Note: poor (1.00–1.75), average (1.76–2.50), fair (2.51–3.25), excellent to good (3.26–4.00)*

As can be seen in the table, the three human raters rated most (i.e. 3 of 4) of the student essay criteria fair, specifically content (M = 2.78; SD = 0.64), language (M = 2.62; SD = 0.41), and mechanics (M = 2.53; SD = 0.38). The fourth criterion, that is organization, was rated average (M = 2.48; SD = 0.48).

On the other hand, the ChatGPT 3.5 ratings of student essays were fair for all four criteria. However, it is worth noting that the ChatGPT average ratings are higher in all four criteria of the rubric compared to the average human ratings. This finding is supported by Parker et al. (2023), who concluded that although ChatGPT demonstrates utility as an automated writing evaluation tool, it can be slightly stricter than human evaluators.

4.2 Reliability of Human and ChatGPT 3.5 Raters in Rating Student Essays

Table 2 shows the level of consistency among the two groups of raters (i.e., human raters versus ChatGPT). The results show the inter-rater agreement along the different criteria used in the essay writing rubric adapted from Carrol and West (1989).

Table 2: Intraclass correlation results indicating reliability of the human and ChatGPT 3.5 essay ratings

Rubric criterion	Reliability coefficient			
	Human raters	Interpretation	ChatGPT 3.5	Interpretation
Content	.609	Moderate	.645	Moderate
Organization	.714	Moderate	.775	Good
Language	.442	Poor	.567	Moderate
Mechanics	.609	Moderate	.593	Moderate
Overall score	**.807**	**Good**	**.724**	**Moderate**

The results show that for the human raters, the consistency was moderate, except for the language criterion, which was interpreted as poor. Overall, the consistency in rating the essays among the human raters was good. This result again indicates the overall good internal consistency of the used rubric in grading student essays.

As to the consistency of the ChatGPT ratings, the data show that ChatGPT obtained a moderate score in three of the four criteria, including content, language, and mechanics, although the organization part of the rubric obtained a good score regarding consistency. The finding is consistent with the study of Latif and Zhai (2023), who indicated that ChatGPT can be fine-tuned for automatic scoring of students' writing work, such as essays.

4.3 Inter-rater Reliability of Human versus ChatGPT 3.5 Ratings of Student Essays

One of the main questions to be addressed in this study was to determine the consistency between the results of the human raters and ChatGPT 3.5. The statistics in this regard are presented in Table 3.

Table 3: Intraclass correlation results indicating inter-rater reliability between human and ChatGPT 3.5 essay ratings

Human raters	ChatGPT 3.5					
	Content	Organiza-tion	Language	Mechanics	Total score	Interpretation
Content	-0.242	–	–	–	–	Poor
Organization	–	-0.182	–	–	–	Poor
Language	–	–	0.56	–	–	Moderate
Mechanics	–	–	–	-0.157	–	Poor
Total score	–	–	–	–	**0.146**	**Poor**

Table 3 shows that, in three of the four criteria for essay writing, the consistency of the human raters and ChatGPT was poor. Mechanics was the only criterion where both ratings were moderately consistent. Overall, the agreement between the two groups of raters was poor (*p* = -0.146). This result indicates that the ratings of the human raters and ChatGPT were not linearly aligned. Hence, in the context of grading students' academic works, such as essays, ChatGPT would still need to improve in its consistency if it were to be used as a replacement for human raters. This finding is similar to the observation of Lund et al. (2023), which reports that although current large language processing models such as ChatGPT are generally quite good, there may still be errors in their interpretations and information generated. The finding further supports the conclusions of Ferrouhi (2023) and Paz et al. (2023), that ChatGPT results may still contain inaccuracy and errors, hence human verification is invaluable.

4.4 Differences between Human and ChatGPT 3.5 Ratings of Student Essays

The test of ICC only shows how the averages of two ratings correlate linearly. Although the human and ChatGPT 3.5 ratings in this study differed, they differed consistently. Hence, further analysis was needed to test for significant differences, as shown in Table 4.

Table 4: Wilcoxon signed-rank test results representing significant differences between the human and ChatGPT 3.5 essay ratings

	Human raters		ChatGPT 3.5		Wilcoxon		
Criterion	X	SD	X	SD	*z*-score	*p*-value	**Interpretation**
Content	2.78	0.64	2.93	0.39	-0.974	0.330	Not significant
Organization	2.48	0.48	2.68	0.50	-0.950	0.342	Not significant
Language	2.62	0.41	2.72	0.41	-0.430	0.667	Not significant
Mechanics	2.53	0.38	2.95	0.70	-2.120	0.034	Significant
Total score	**11.28**	**1.58**	**11.24**	**1.91**	**-1.530**	**0.126**	**Not significant**

**Significant at a ≤ 0.05*

As can be seen in the table, the mean scores for the ratings of student essays by ChatGPT are slightly higher in all four criteria compared to that of the human raters. However, further analysis using the Wilcoxon signed-rank test showed no significant difference in both ratings for three of the criteria, namely content (z = -0.974; p = 0.330), organization (z = -0.950; p = 0.342), and language (z = -0.430; p = 0.667). The only criterion for which the rating between the human raters and ChatGPT 3.5 significantly differed was mechanics (z = -2.120; p = 0.034). In terms of the overall rating, the results also show no significant difference (z = -1.530; p = 0.126).

Generative AIs, such as ChatGPT, pose many potentials. The findings of this study confirm that ChatGPT could be useful to teachers as it may help in rating students' work, which will free teachers from many teaching-related tasks to focus on instruction. This study has shown that, with proper prompts from the teacher, ChatGPT can be used to rate students' written work. The data show that ChatGPT tends to give a somewhat higher score than human raters when rating student essays, although such a difference is not statistically significant. Finally, the results of the comparative analysis show a poor to moderate consistency of ChatGPT rating with that of human raters. Hence, ChatGPT should be used with caution, a finding which is consistent with the conclusion of some previous studies (Bitzenbauer, 2023; Ferrouhi, 2023; Kasneci et al., 2023; Lo, 2023).

5. Conclusion

This study was conducted to determine the consistency of ChatGPT with human raters in terms of rating student essays to shed light on the potential integration of ChatGPT into classrooms. The results show that among the human raters, the consistency was good, indicating the reliability of the used rubric. As to the consistency of the ratings by ChatGPT 3.5, the inter-rater coefficient was moderate. In terms of the consistency of ratings between the human raters and ChatGPT, the results reveal that there was a poor consistency along three of the four criteria, namely content, organization, and language. Overall, the consistency of the two groups of raters was poor, suggesting that the ratings of human raters and ChatGPT are not linearly related. Further analysis of the results showed that the mean scores for ChatGPT ratings of student essays were slightly higher in all four criteria compared to those for the human raters. However, the difference was found to be not statistically significant. The study concludes that ChatGPT 3.5 in its current version requires human assistance to verify the accuracy of its provided assessment, especially in the context of rating student essays.

6. Recommendations

Some important practical implications for classroom and research purposes can be drawn based on the results of the study.

For teachers, ChatGPT should be used cautiously in rating students' written works, as scores are important factors for determining students' passing or failing grades in schools or for identifying top students in class. While teachers may use ChatGPT as a potentially useful assistive tool in scoring student works, it remains their ethical responsibility to validate the data generated by this generative AI tool, ensuring that the accuracy of information is not compromised.

For future research, it would be useful to extend the current findings by examining the consistency of ChatGPT in rating involving more student participants and using other types of student written works, such as critique papers and reflective essays. This study made use of ChatGPT version 3.5, which is the most advanced free version of OpenAI's generative AI. However, other studies may investigate using more advanced versions of generative AI tools, such as ChatGPT 4 and Google's Bard.

7. Study Limitations

This study was not without limitations. One of the limitations is the limited sample, as the participants were selected from only one group of college students. Including a more diverse group of students might have provided different interesting results. In addition, the study used student essays only to establish inter-rater consistency, rather than varied student works. Other student works, such as critique papers, could have yielded interesting insights in establishing the inter-rater reliability of rating.

Acknowledgement

The authors wish to thank the students who participated in this study and all those who contributed directly or indirectly to its completion.

Declaration

The authors declare no conflict of interest in the conduct and publication of this study.

9. References

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology, 15*(3), Article 429. https://doi.org/10.30935/cedtech/13152

Alrishan, A. M. (2023) Determinants of intention to use ChatGPT for professional development among Omani EFL pre-service teachers. *International Journal of Learning, Teaching and Educational Research*, *22*(12), 187–209. https://doi.org/10.26803/ijlter.22.12.10

Biswas, S. (2023, June 4). Role of ChatGPT in education. *SSRN.* https://ssrn.com/abstract=4369981

Bitzenbauer, P. (2023). ChatGPT in physics education: A pilot study on easy-to-implement activities. *Contemporary Educational Technology*, *15*(3), ep430. https://doi.org/10.30935/cedtech/13176

Buchholz, K. (2023, July 7). Threads shoots past one million user mark at lightning speed. *Statistica*. https://www.statista.com/chart/29174/time-to-one-million-users/

Dergaa, I., Chamari, K., Zmijewski, P., & Saad, H. B. (2023). From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. *Biology of Sport*, *40*(2), 615–622. https://doi.org/10.5114/biolsport.2023.125623

Dikli, S., & Bleyle, S. (2014) Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessment Writing, 22*(1), 1−17. https://doi.org/10.1016/j.asw.2014.03.006

Ferrouhi, E. M. (2023). Evaluating the accuracy of ChatGPT in scientific writing. *Research Square* [preprint]. https://doi.org/10.21203/rs.3.rs-2899056/v1

Firat, M. (2023, January 12). *How ChatGPT can transform autodidactic experiences and open education?* https://doi.org/10.31219/osf.io/9ge8m

Fuchs, K. (2023). Exploring the opportunities and challenges of NLP models in higher education: Is Chat GPT a blessing or a curse? *Frontiers in Education, 8*, Article 1166682. https://doi.org/10.3389/feduc.2023.1166682

Gill, S. S., Xu, M., Patros, P., Wu, H., Kaur, R., Kaur, K., Fuller, S., Singh, M., Arora, P., Parlikad, A. K., Stankovski, V., Abraham, A., Ghosh, S. K., Lutfiyya, H., Kanhere, S. S., Bahsoon, R., Rana, O., Dustdar, S., Sakellariou, R., Uhlig, S., & Buyya, R. (2024). Transformative effects of ChatGPT on modern education: Emerging era of AI chatbots. *Internet of Things and Cyber–Physical Systems*, *4*, 19−23. https://doi.org/10.1016/j.iotcps.2023.06.002

Harunasari, S. Y. (2023). Examining the effectiveness of AI-integrated approach in EFL writing: A case of ChatGPT. *International Journal of Progressive Sciences and Technology (IJPSAT)*, *39*(2), 357–368. https://ijpsat.org/index.php/ijpsat/article/download/5516/3447

Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology*, *15*(4), ep464. https://doi.org/10.30935/cedtech/13605

Iqbal, N., Ahmad, H., & Azhar, K. (2023). Exploring teachers' attitudes towards using ChatGPT. *Global Journal for Management and Administrative Sciences, 3*(4), 97–111. https://doi.org/10.46568/gjmas.v3i4.163

Kasneci, E., Sebler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Gunnemann, S., Hullermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kopp, W., & Thomsen, B. (2023, May 1). How AI can accelerate student's holistic development and make teaching more fulfilling. *World Economic Forum*. https://www.weforum.org/agenda/2023/05/ai-accelerate-students-holistic-development-teaching-fulfilling/

Laschinger, H. K. (1992). Intraclass correlations as estimates of interrater reliability in nursing research. *Western Journal of Nursing Research*, *14*(2), 246–251. https://journals.sagepub.com/doi/pdf/10.1177/019394599201400213

Latif, E., & Zhai, X. (2023) Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, Article 100210. https://doi.org/10.1016/j.caeai.2024.100210

Li, L., Ma, Z., Fan, L., Lee, S., Yu, H., & Hemphill, L. (2023). ChatGPT in education: A discourse analysis of worries and concerns on social media. *Cornell University.* https://doi.org/10.48550/arXiv.2305.02201

Liu, M., Xu, W., Ran, Q., & Li, Y. (2015). Using natural language processing technology to analyze teachers' written feedback on Chinese students' English essays. *International Journal of Learning, Teaching and Educational Research, 11*(1), 1–11. https://ijlter.net/index.php/ijlter/article/view/1087

Lingard, L. (2023). Writing with ChatGPT: An illustration of its capacity, limitations & implications for academic writers. *Perspective on Medical Education, 12*(1), 261–270. https://doi.org/10.5334%2Fpe.1072

Livberber, T. (2023). Toward non-human-centered design: Designing an academic article with ChatGPT. *Profesional de la Información*, *32*(5), 1–19. https://doi.org/10.3145/epi.2023.sep.12

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), Article 410. https://doi.org/10.3390/educsci1304041

Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology, 74*(5), 570–581. https://doi.org/10.1002/asi.24750

McNamara, D., Crossley, S., Roscoe, R., Allen, L., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing, 23*, 35–39. https://researchlanglit.gsu.edu/files/2016/08/272.pdf

Mondal, H., & Mondal, S. (2023). ChatGPT in academic writing: Maximizing its benefits and minimizing the risks. *Indian Journal of Ophthalmology*, *71*(12), 3600–3606. https://doi.org/10.4103/IJO.IJO_718_23

Mondal, H., Marndi, G., Behera, J. K., & Mondal, S. (2023). ChatGPT for teachers: Practical examples for utilizing artificial intelligence for educational purposes. *Indian Journal of Vascular and Endovascular Surgery, 10,* 200–205. https://doi.org/10.4103/ijves.ijves_37_23

Morozov, E. (2023, July 5). The true threat of artificial intelligence. *International New York Times.* https://link.gale.com/apps/doc/A755774581/AONE?u=anon~c95b5477&sid=googleScholar&xid=52964d0e

Parker, J., Becker, K., & Corroca, C. (2023). ChatGPT for automated writing evaluation in scholarly writing instruction. *Journal of Nursing Education, 62*(12), 721–727. https://doi.org/10.3928/01484834-20231006-02

Paz, M. A., Turner, K., & Racila, E. (2023). Evaluating the performance of ChatGPT in writing autopsy clinicopathological correlations. *American Journal of Clinical Pathology, 160*(1), S125. https://doi.org/10.1093/ajcp/aqad150.272

Rahman, M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences, 13*(9), Article 5783. https://doi.org/10.3390/app13095783

Sabzalieva, E., & Valentini, A. (2023). *ChatGPT and artificial intelligence in higher education: Quick start guide*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000385146

Sharma, S., & Yadav, R. (2023). Chat GPT: A technological remedy or challenge for education system. *Global Journal of Enterprise Information System*, *14*(4), 46–51. https://www.gjeis.com/index.php/GJEIS/article/view/698

Sharples, M. (2022). Automated essay writing: An AIED opinion. *International Journal of Artificial Intelligence in Education, 32,* 1119–1126. https://doi.org/10.1007/s40593-022-00300-7

Siedlecki, S. (2020) Understanding descriptive research designs and methods. *Clinical Nurse Specialist, 34*(1), 8–12. https://doi.org/10.1097/NUR.0000000000000493

Trust, T., Whalen, J., & Mouza, C. (2023). Editorial: ChatGPT: Challenges, opportunities, and implications for teacher education. *Contemporary Issues in Technology and Teacher Education, 23*(1), 1–23. https://www.learntechlib.org/primary/p/222408/

Tuyen, T., Osman, S. B., Ahmad, N. S. B., & Dan, T. C. (2019). *Developing and validating scoring rubrics for the assessment of research papers writing ability of EFL/ESL undergraduate students: The effects of research papers writing intervention program using process genre model of research paper writing.* https://www.semanticscholar.org/paper/Developing-and-Validating-Scoring-Rubrics-for-the-Tuyen-Osman/e86657da7ec761a7d403b4f8d85cc8bf19f99923

Vargas-Murillo, A. R., Pari-Bedoya, I., & Guevara-Soto, F. (2023). Challenges and opportunities of AI-assisted learning: A systematic literature review on the impact of ChatGPT usage in higher education. *International Journal of Learning, Teaching and Education Research, 22*(7), 122–135. https://doi.org/10.26803/ijlter.22.7.7

Waltzer, T., Cox, R., & Heyman, G. (2023). Testing the ability of teachers and students to differentiate between essays generated by ChatGPT and high school students. *Human Behavior and Emerging Technologies*, Article 1923981. https://doi.org/10.1155/2023/1923981