Using ChatGPT-4 to grade open question exams

Hani Alers[1], Aleksandra Malinowska[2], Gregory Meghoe[1] and Enso Apfel[1]

[1] The Hague University of Applied Sciences, Zoetermeer, The Netherlands
[2] University of California, Santa Barbara, California, USA
HAL@HHS.NL

Abstract. This research investigates the potential and challenges of using artificial intelligence, specifically the ChatGPT-4 model developed by OpenAI, in grading and providing feedback in an educational setting. By comparing the grading of a human lecturer and ChatGPT-4 in an experiment with 105 students, our study found a strong positive correlation between the scores given by both, despite some mismatches. In addition, we observed that ChatGPT-4's feedback was effectively personalized and understandable for students, contributing to their learning experience. While our findings suggest that AI technologies like ChatGPT-4 can significantly speed up the grading process and enhance feedback provision, the implementation of these systems should be thoughtfully considered. With further research and development, AI can potentially become a valuable tool to support teaching and learning in education.

Keywords: AI in Education; ChatGPT-4 Grading; Automated Feedback Systems; AI vs. Human Assessment; Educational Technology Advancements.

1 Introduction

In the rapidly evolving realm of education, artificial intelligence, specifically Large Language Models (LLM) like ChatGPT, has emerged as a potentially transformative tool. While the integration of these technologies holds considerable promise, it equally raises concerns about their responsible and ethical usage.

Previous studies have demonstrated the myriad of advantages LLMs like ChatGPT offer in education. For instance, they can enhance student productivity, optimize time efficiency, assist with inquiries, and even foster collaboration among learners [2, 3, 4, 8]. In a survey, approximately 96% of students expressed interest in using ChatGPT, with 83% inclined to use it more frequently for academic purposes [5].

However, the adoption of such technology is not without reservations. The reliability of the information, for one, requires a critical evaluation. Despite its capabilities, ChatGPT cannot wholly substitute human intelligence, and students must possess a solid background knowledge to discern the accuracy of its outputs

[5, 7]. Ethical implications, particularly concerning privacy and bias, remain pertinent issues [1].
Building on these findings and concerns, this study raises the question: "How close can ChatGPT get to grading and providing feedback compared to a human teacher?" This inquiry stems from a tangible challenge in the educational sector: the labor-intensive nature of grading, especially for open-ended questions, and the often daunting task of securing competent graders.
To answer this research question, we plan to have ChatGPT-4 grade and provide feedback on a test that has already been made and evaluated. We will use a script specifically designed for ChatGPT-4 to grade the tests. This script includes the text describing the case study used as the subject for the exam questions, the questions themselves, the grading form, the correct answer key, the student's answers, and prompts to simulate the grading process as accurately as possible.

2 Methodology

2.1 Description of the exam used in the experiment

The core of our research experiment hinged on an existing exam from The Hague University of Applied Sciences. This exam was given to 108 second-year students from the HBO-ICT program during the midst of the Covid-19 pandemic, necessitating a digital format for completion. A team of three lecturers from the institution had graded this exam back in 2021, with each instructor being responsible for their designated set of questions, ensuring a uniform grading standard for all student submissions. The exam's questions centered around an article that detailed a research project case study. Students were prompted to answer questions about the research case study asking to identify aspects like the nature of the research, the data gathering methodology used, and so on.
To maintain the integrity and ethics of our study, we ensured that all identifiable student information was anonymized. Furthermore, neither the students nor the faculty were aware at the time of the exam that these tests would later become part of a research endeavor. This unsuspecting approach strengthens the reliability of our findings since the original exam was designed and graded without any prospective research bias.
The data used for the study included the case study that was given to the students, the exam questions about that case study, the anonymized answers from 108 students, and the exam's answer key used for grading. In total, there are eleven points to earn for this exam, and the test contains three different main questions, each with several sub-questions. Question 1 contains four sub-questions, each worth one point, question 2 has three sub-questions, each also worth one point, and question 3 has three sub-questions, where the first sub-question is worth two

points, and the rest one point each. The questions asked in the exam consistently relate to the text of the case study, and the student must explain their answer in coherent flowing text. Although the answers in the answer key contain a keyword, the student also needs to provide an explanation incorporating this specific keyword. For question 3, the student also needs to quote a piece of text from the case study. Below you can see the first question from the exam (the question is translated from Dutch to English):

- ***Q1 pt. a. Explain in a coherent text whether the research is fundamental or applied.***
- ***Q1 pt. b. Explain in a coherent text what type of objective this research has.***
- ***Q1 pt. c. Explain in a coherent text what the nature of the research is.***
- ***Q1 pt. d. Explain in a coherent text whether this research is qualitative or quantitative.***

2.2 Grading the exam using ChatGPT-4

For the research experiment, a script has been compiled to instruct ChatGPT-4. The script contains information from the exam and prompts to simulate the grading and assessment process of a human teacher. The script that is inputted into ChatGPT-4 includes the text of the provided case study, the exam questions about the case study, the student's answers to the exam questions, and the answer key to the exam questions containing key word responses. The script also contained specific prompts engineered to facilitate the best performance from ChatGPT-4. Below is a an example prompt in the script:

- `Check the student's answers and compare them with those of the teacher\`
- `When a question is answered wrong, provide feedback based on the correct answer and text\`
- `When a question is answered correctly, print out 'correct' and do not give any feedback\`

The script created for ChatGPT-4 follows the latest guidelines provided by OpenAI (the company behind ChatGPT) at the time of writing this article [9]. The script used for the grading was optimized for best performance using a training set of 3 students. To avoid overfitting, these 3 students were not included in the dataset, leaving 105 participants for the analysis. Each time the script is fed into ChatGPT-4 containing the answers from one student. The script functions as a foundation for getting the information out of the exams. It remains unchanged, only the answers from the students vary each time. Furthermore, the output obtained from ChatGPT-4 is manually copied to a separate spreadsheet. Within this spreadsheet, the output which contains the questions, student answers,

teacher answers, points per question, and feedback is copied for each of the 105 students. The data will then be analyzed using this spreadsheet.

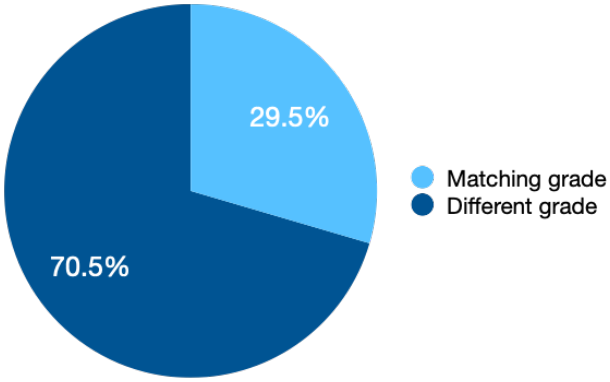

Fig1. Comparison of grade for ChatGPT vs Human

3 Results and analysis

3.1 ChatGPT grading compared to a lecturer

The study began with a detailed look at the total scores given by a human teacher and ChatGPT-4. Figure 1 shows that for 74 students (representing 70.5% of the cases), the total scores from the human lecturer and that ChatGPT-4 did not match. Still, a Pearson correlation test showed that there was a positive correlation between the two variables with $r = 0.878$, $n = 105$, $p < 0.001$. This means that while the exact scores did not always match, there was a strong relationship in the scoring trends. Meaning, when the human teacher gave high grades, ChatGPT-4 also tended to give high grades, and the other way around, as shown in Figure 2.
Despite the strong correlation, a paired sample t-test showed that there was still a significant difference in the scores for lecturer grades ($M=5.073$, $SD=2.567$) and GPT grades ($M=5.321$, $SD=2.512$) conditions; $t(108)=-2.056$, $p = 0.042$, Cohen's $D = 0.197$. Since Cohen's D is smaller than 0.2, we can conclude that although the difference is significant, the effect size is still quite small. The means are plotted in Figure 3, with error bars representing the standard error of the mean. The figure clearly shows that the outcomes are quite similar.

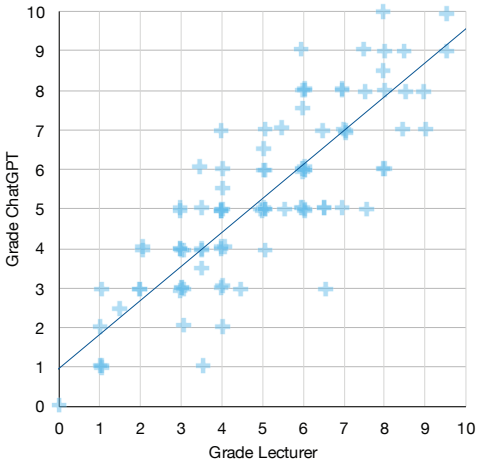

Fig2. Scatter plot of final grade given by ChatGPT-4 and human lecturer. Grades have increments of 0.5 points, however the dots in the graph have been shifted for better visibility

In our analysis of the grading outcomes, a noteworthy observation emerged concerning the consistency between human and ChatGPT-4 grading. Out of the 105 students in the study, grading alignment – where both the human teacher and ChatGPT-4 either passed or failed a student – was observed for 86 students. This signifies a concurrence rate of approximately 81.9%. However, a discrepancy was evident in the case of 19 students, wherein they received a passing grade under one grading condition (either human or ChatGPT-4) but not the other. This mismatch underscores the need for a deeper examination of the specific criteria

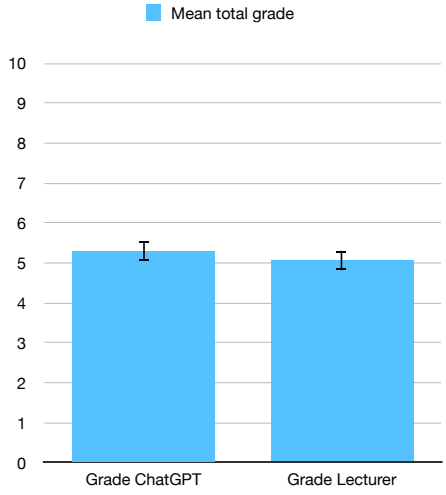

Fig3. Mean total grade, error bars represent standard error of the mean

and nuances that might account for such variations between human and AI grading approaches.

The final part of our analysis looked at differences in scoring per question. Figure 4 shows the match in scores for individual questions. The figure shows that the match in scores for Q1 and Q2 is much higher than for Q3. This can be due to the complexity of the Q3. Answers to Q1 and Q2 were simpler, often including just one keyword or concept. In contrast, Q3 required a specific piece of text from the case study in the answers, leading to longer and more complex answers from the students.

In conclusion, while there are similarities in the total scores awarded by the human teacher and ChatGPT-4, discrepancies emerged in the pass rates that cannot be overlooked. The most pronounced variations appeared to be in questions that necessitated longer and more intricate responses. The alignment in scores for a majority of the students suggests that the correlation between the grades assigned by the human teacher and ChatGPT-4 wasn't merely coincidental. However, with a non-trivial number of students receiving divergent pass-fail outcomes between the two grading entities, it becomes evident that while AI, like ChatGPT-4, displays promise in evaluating student performances, there are nuances and complexities that might not yet fully align with human judgment.

3.2 Feedback from ChatGPT

Our research into the feedback function of ChatGPT-4 revealed that this system effectively uses its extensive knowledge database to provide accurate, relevant, and understandable feedback. The AI's ability to directly refer to the relevant study texts was particularly useful in helping students understand why their answers were incorrect or incomplete. A sample of this feedback can be seen in Table 1 below.

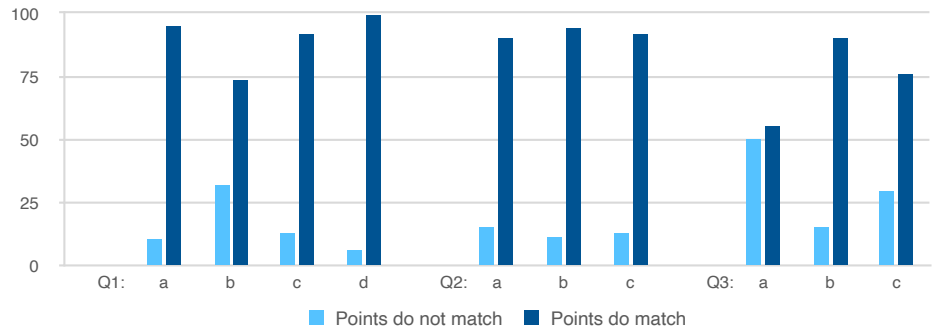

Fig4. Number of (mis)matching grades per sub question of the exam when graded by ChatGPT vs lecturers

Table 1. Example of feedback by ChatGPT-4

Question	Answer student	Correct answer	Feedback ChatGPT-4
Explain in a coherent text what the nature of the research is.	The nature of this research is descriptive. The situation is explained and mapped out, and the research is quantitative.	Evaluative	The research is evaluative, as it tests the function of the brachytherapy cement in practice, not just describes it.
Explain in a coherent text what type of goal this research has.	Acquiring knowledge to share with the world. This research has lasting value; the researchers aim to gain knowledge that can also be applied to, for example, tumor patients.	Solving a problem	The research seems to be aimed at solving a specific problem (finding a safer alternative to conventional radiation therapy) rather than just acquiring knowledge.

For this exam, the students did not receive feedback from the teachers who marked the exam. However, during the grading of the tests, we asked ChatGPT-4 if it could provide feedback when a question is answered incorrectly.

A key observation pertained to ChatGPT-4's adeptness in linguistic adaptability and personalization. The AI demonstrated an ability to adjust its feedback language to match that of the student, enhancing the clarity and accessibility of its responses. Whether the student employed intricate terms or opted for more straightforward language, ChatGPT-4 tailored its feedback, not just in linguistic complexity but also in alignment with the unique learning needs and styles of each student. This adaptive style leads to a more effective and efficient learning experience.

In conclusion, the results of our research demonstrate that ChatGPT-4 is a valuable tool for providing high-quality, personalized feedback that can enhance students' learning and understanding.

3.3 Advantages of using ChatGPT for grading

The use of AI technology, like ChatGPT-4, in schools can bring big benefits, especially when it comes to grading and giving feedback. One of the best parts of using this method is how fast it is. In our study, we found that grading a test took about a minute. This minute included putting students' answers into our script, letting ChatGPT-4 grade the test, and then putting the grades from ChatGPT-4 into the Excel file. In total, it took ChatGPT-4 about 105 minutes to grade all student tests. When we compare this to a team of three teachers who took a full week, or about 7200 minutes, we see a huge difference. In simple terms, using ChatGPT-4 took only about 1.5% of the time that the lecturers needed. Keep in

mind that ChatGPT-4 also provided feedback when the answers were incorrect while the human graders did not.

On the other hand, several limitations warrant attention. The script used in this study was tailored for a specific test, and applying it to different tests might not produce accurate grades. This necessitates the creation of a unique script for each test. While this preliminary setup can be time-consuming, the accelerated grading process offered by ChatGPT-4 is likely to compensate for the initial time investment.

The specific AI technology utilized plays a pivotal role. In this study, we chose ChatGPT-4 after ChatGPT-3.5 did not meet our expectations. Although ChatGPT-3.5 responded as if it was performing the task correctly, the grades and feedback it provided were completely wrong. Even with ChatGPT-4, we encountered challenges including a restriction to 25 prompts followed by a 3-hour wait, and occasional grading inconsistencies. Moreover, during high-demand periods, ChatGPT's response quality diminished. These limitations prevented us from grading all tests simultaneously, extending the grading duration. This extra time was not included in the 105 minutes calculation as utilizing the ChatGPT API can bypass the question time limit. Therefore this article focuses purely on the capabilities of the AI driving ChatGPT.

4 Discussion

Firstly, the results suggest that ChatGPT-4, although not perfect, has significant potential as a tool for grading and giving feedback in an educational setting, which supports what is already found in the literature [6]. The strong correlation between the scores given by ChatGPT-4 and a human lecturer, despite some differences, points to the AI's potential to maintain grading performance similar to human educators. However, it is important to note that the current capacity of AI technologies like ChatGPT-4 might not cover the full breadth of assessing student understanding, especially for more complex, text-based answers.

Our research also showcased the potential of ChatGPT-4 to provide relevant and understandable feedback to students. We found the AI's ability to adjust its language use and personalize feedback based on student answers particularly promising. This can significantly contribute to an effective and efficient learning process by making feedback more accessible and personalized.

However, applying AI like ChatGPT-4 in educational settings has certain limitations. A major challenge is the current need for a unique script or program for each specific test, meaning that the system lacks the ability to generalize across different types of exams or assignments. While ChatGPT-4 offers impressive time-saving performance, the preparatory work required for grading each individual exam is something to keep in mind.

Moreover, while ChatGPT-4's grading speed dramatically outpaces that of a team of human lecturers, there were some challenges with the grading process itself, which could sometimes be irregular or require breaks between grading batches. These can potentially contribute to the time and resource costs of implementing this AI solution in a practical environment. Therefore, our research highlights that while AI technologies like ChatGPT-4 offer promising benefits for education, their current limitations and potential pitfalls need careful consideration. It's crucial to remember that AI, as it stands now, cannot replace the unique and nuanced understanding that human lecturers bring to the grading and feedback process [5, 7].

Looking ahead, our research paves the way for further exploration into the applications of AI in education. Future studies can focus on refining the interaction scripts with ChatGPT-4, testing the AI's grading and feedback capabilities for different types of assessments or levels of education, and exploring how best to integrate AI support with human teaching methods. It would also be useful to investigate the broader implications of AI-supported education for the teaching profession and students' learning experiences. With continued research and improvements, AI technology could potentially become a valuable addition to human teaching, enhancing the educational process.

5 Conclusion

This study provided important insights into the application of AI technology, specifically ChatGPT-4, in an educational setting, focusing on the grading of student work and the provision of feedback. In comparing the total scores given by a human lecturer and ChatGPT-4, it was found that while they did not match in 70,5% of the cases, there was a strong positive correlation in the scoring trends. Nevertheless, pass rates were significantly affected, indicating that using AI to grade exams still needs further development.

The analysis of ChatGPT-4's feedback demonstrated effective use of its extensive knowledge database to deliver accurate and relevant information in an understandable format. Particularly, the AI's adaptive language ability and personalization catered to the unique learning needs of students, thereby improving their learning experience and understanding.

Despite these benefits, it is crucial to acknowledge the challenges that come with employing AI in educational contexts. Although AI grading saves significant time, the complexity of adapting the system to different tests and occasional discrepancies in grading highlights the need for continuous refinement of such technologies.

In conclusion, our findings suggest that while ChatGPT-4 and similar AI technologies hold substantial promise for educational purposes, their

implementation should be thoughtfully considered. With continuous research and improvements, the use of AI can become a valuable addition to our educational systems, augmenting lecturers' efforts and enhancing students' learning experiences.

References

1. Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? Library Hi Tech News. Available at SSRN: https://ssrn.com/abstract=4333415 or http://dx.doi.org/10.2139/ssrn.4333415
2. Belhaj, N., Hamdane, A., Chaoui, N. E. H., Chaoui, H., & Bekkali, M. E. (2021). Engaging students to fill surveys using chatbots: University case study. Indonesian Journal of Electrical Engineering and Computer Science, 24(1), 473. https://doi.org/10.11591/ijeecs.v24.i1.pp473-483
3. Fauzi, F., Tuhuteru, L., Sampe, F., Ausat, A. M. A., & Hatta, H. R. (2023). Analysing the Role of ChatGPT in Improving Student Productivity in Higher Education. Journal on Education, 5(4), 14886–14891. https://doi.org/10.31004/joe.v5i4.2563
4. Clarizia, F., Colace, F., Lombardi, M., Pascale, F., & Santaniello, D. (2018). Chatbot: An education support system for student. In Cyberspace Safety and Security: 10th International Symposium, CSS 2018, Amalfi, Italy, October 29–31, 2018, Proceedings 10 (pp. 291-302). Springer International Publishing.
5. Shoufan, A. (2023). Exploring Students' Perceptions of ChatGPT: Thematic Analysis and Follow-Up Survey. IEEE Access, 11, 38805-38818.
6. Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2022). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health, 2.
7. Jalil, S., Rafi, S., LaToza, T.D., Moran, K., & Lam, W. (2023). ChatGPT and Software Testing Education: Promises & Perils. 2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), 4130-4137.
8. Tack, A., & Piech, C. (2022). The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues. ArXiv. Retrieved from https://arxiv.org/abs/2205.07540
9. OpenAI Platform. (2023). Explore developer resources, tutorials, API docs, and dynamic examples to get the most out of OpenAI's platform. Retrieved from https://platform.openai.com/examples