Received: 3 June 2024 | Accepted: 28 August 2024

DOI: 10.1002/berj.4069

BERJ BERA

ORIGINAL ARTICLE

Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT

Jonas Flodén

Department of Business Administration, School of Business, Economics and Law, University of Gothenburg, Gothenburg, Sweden

Correspondence
Jonas Flodén, Department of Business Administration, School of Business, Economics and Law, University of Gothenburg, Box 610, Gothenburg 405 30, Sweden.
Email: jonas.floden@handels.gu.se

Abstract

This study compares how the generative AI (GenAI) large language model (LLM) ChatGPT performs in grading university exams compared to human teachers. Aspects investigated include consistency, large discrepancies and length of answer. Implications for higher education, including the role of teachers and ethics, are also discussed. Three Master's-level exams were scored using ChatGPT 3.5, and the results were compared with the teachers' scoring and the grading teachers were interviewed. In total, 463 exam responses were graded. With each response being graded at least three times, a total of 1389 gradings were conducted. For the final exam scores, 70% of ChatGPT's gradings were within 10% of the teachers' gradings and 31% within 5%. ChatGPT tended to give marginally higher scores. The agreement on grades is 30%, but 45% of the exams received an adjacent grade. On individual questions, ChatGPT is more inclined to avoid very high or very low scores. ChatGPT struggles to correctly score questions closely related to the course lectures but performs better on more general questions. The AI can generate plausible scores on university exams that, at first glance, look similar to a human grader. There are differences but it is not unlikely that two different human graders could result in similar discrepancies. During the interviews, teachers expressed their surprise at how well ChatGPT's grading matched their own. Increased use of AI can lead to ethical challenges as exams are entrusted to a machine whose decision-making criteria are not fully understood, especially concerning potential bias in training data.

KEYWORDS
AI, exam, grading, higher education

Key insights

What is the main issue that the paper addresses?

The introduction of artificial intelligence (AI) and large language models (LLMs) allows exams to be scored automatically without prior training of the model. This paper addresses the question of how the automatic scoring performs in comparison with human teachers in higher education.

What are the main insights that the paper provides?

Without detailed grading instructions, the model can generate scores on university exams that, at first glance, look similar to those of a human grader. However, it tends to produce more medium scores and fewer extreme ones compared to humans, resulting in grade differences in 70% of cases.

INTRODUCTION

Artificial intelligence (AI) is rapidly making an impact on the higher education sector. Many potential applications have been suggested for artificial intelligence in education (AIEd), ranging from AI-assisted student training to assessment and tutoring (Bond et al., 2024; Crompton & Burke, 2023; Holmes & Tuomi, 2022; Miao et al., 2021), but until late 2022 the application of AI technology was not widespread. However, the release of the generative AI (GenAI) chat robot ChatGPT-3 in December 2022 changed this situation entirely and shocked the education sector. Multiple examples were made public of students submitting their assignments to ChatGPT and receiving high marks (Farazouli et al., 2023). Studies showed that ChatGPT was able to pass exams in, for example, physics (Yeadon et al., 2023), medicine (Kung et al., 2023), law (Choi et al., 2023), surgery (Gencer & Aydin, 2023) and operations research (Terwiesch, 2023). As the AI generated a unique text each time, existing plagiarism software was unable to detect it (Yeadon et al., 2023) and teachers had difficulty separating a ChatGPT-written answer from a real student answer (Farazouli et al., 2023).

Just as ChatGPT is able to generate answers to exam questions, it can also be asked to grade exams. Grading is a high-stakes task as an incorrect grading decision might have severely negative effects for students or for the general public, who might get people working in jobs they are not qualified for. The assessment must be fair and accountability high, as the consequences of an incorrect assessment are high (Stobart & Eggen, 2012). From both an ethical and a legal perspective, students have the right to be given a fair evaluation (Russell & Airasian, 2012). It is therefore necessary that exam grading be performed fairly. At the same time, grading is often considered a boring, tedious and time-consuming task by teachers. Just as students will be tempted to use ChatGPT to write their assignments, teachers will be tempted to use ChatGPT for grading. Similarly, given the rapid development in the AI field, we are likely to be seeing new AI tools emerging designed for grading. It is therefore of high importance to establish the extent to which AI gradings are trustworthy.

A remarkable aspect of large language models (LLMs), such as ChatGPT, is that they do not require any specific training or instructions to be able to perform grading. LLMs allow for a zero-shot strategy (Kojima et al., 2022) in which a user can just ask them to grade any question in any subject, without providing any instructions for how the grading should be done and what criteria should be considered. Previous systems for automated scoring have required extensive training on already scored exam questions to create a scoring model (Burrows et al., 2015; Zupanc & Bosnić, 2015). Although it is possible to provide more detailed instructions and grading rubrics to ChatGPT as well, it is likely that many users will be tempted by the ease of use and employ ChatGPT as a zero-shot grader. This raises the question of how good the scoring is compared to a human grader when used as a zero-shot grader. The purpose of this study was to investigate if the scoring by ChatGPT as a zero-shot grader on exams in higher education was comparable to the scoring of a human grader and, if not, to determine the ways in which they differ.

This paper investigates this point by comparing the scoring of 463 ChatGPT 3.5-scored exam questions with the score from human graders. Further, explorative interviews were performed with the grading teachers. The paper discusses how much agreement there was on exam grades and scoring. In particular, exam questions with a sizable discrepancy were investigated. It also studies AI's grading consistency, impact of length of answer and spread of scoring.

The paper is structured as follows. The following section provides a literature overview of assessment using AI, followed by a description of the methods used. The results are then presented, followed by a discussion and conclusions.

AUTOMATED SCORING USING AI

The assessment of student performance is a natural part of higher education. This assessment can take many forms (Russell & Airasian, 2012), with one of the most common forms of evaluation being the traditional written exam. The idea of using technology to grade exams is not new. Multiple-choice questions have long been subjected to automated scoring, for example, especially in large national exams such as SATs in the United States. However, it remains more challenging to automatically grade open essay-style questions, which have been identified as one possible application of AI in higher education (Celik et al., 2022; Gardner et al., 2021; Xu & Ouyang, 2022). Several AI approaches have been suggested. See, for example, Borade and Netak (2021), Uto (2021) and Bai and Stede (2022) for reviews. Ramesh and Sanampudi (2022) have conducted a comprehensive review of essay-scoring systems, identifying 62 previously studied and concluding that significant challenges still remain, with similar results being found in a review by Ifenthaler (2022). Burrows et al. (2015), in an earlier review, identified 35 systems for short-answer scoring, while attempts were made as early as 1966 (Page, 1966) to create automated scoring systems. Most approaches have been research projects not available to the general public and requiring technical skills to operate, although commercial systems do exist (see Zupanc and Bosnić, 2015 for an overview). However, the use of these earlier systems has been limited (Holmes & Tuomi, 2022). A general issue is that they have difficulties giving the depth and accuracy of analysis in their feedback that a teacher is capable of (Holmes & Tuomi, 2022). Instead, they focus more on surface features of the writing, linguistics and text production skills (Deane, 2013). Ke and Ng (2019) suggest that existing automated scoring systems to a great extent have managed to address dimensions such as grammar, word use and sentence structure, but that challenges still remain with dimensions such as coherence, clarity and persuasiveness of arguments. Early systems, such as Project Easy Grader (PEG), developed by Page in 1966, used statistics and correlations to predict the quality of the text. These systems were criticised for focusing on superficial text features, such as length and word order, that could be correlated with a human grader (Dikli, 2006; Landauer et al., 2000). Later systems, such as Intelligent Essay Assessor (IEA), developed in the late 1990s, utilised machine learning and more advanced statistical methods to compare student texts with high-quality texts. Although these systems were able to consider the semantic content and understanding of the text to a greater extent, they were also criticised for focusing on surface features (Dikli, 2006). With the further development of AI, natural language processing (NLP) started to be applied in systems such as E-rater. This allowed a broader range of features to be assessed and provided more detailed feedback on the texts. Ramesh and Sanampudi (2022) provide an overview of machine learning and NLP models concluding that, despite numerous techniques evaluated, challenges still remain with regard to considering cohesion, coherence, completeness and feedback.

Studies of previous automated grading systems showed that such systems could provide scoring comparable to that of human graders from a numerical perspective (Ifenthaler, 2022; Powers et al., 2000; Zawacki-Richter et al., 2019; Zupanc & Bosnić, 2015). However, scoring validity also implies an understanding of the scope and purpose of the exam. Williamson et al. (2012) suggest a framework for using automated scoring that underlines the importance of a good conceptual fit between the scoring system and the task at hand, to ensure that the system measures the intended construct and that the appropriate data are used to train the model. Other studies further highlight the need to consider that the scores can be explained by the construct, that the answer aligns with other independent performance measures and shows a similarity in scoring between different tasks (Elliot & Williamson, 2013; Huawei & Aryadoust, 2023). The demands on an automated system should also be higher if it is used for a high-stakes decision, such as a final grade, than if it is used for low-stakes decisions, such as pre-screening (Williamson et al., 2012). It has also been argued that a machine cannot truly understand a text (Attali, 2013) and that a good text should show aspects such as curiosity, creativity and flexibility, aspects which pose a challenge for automated scoring systems (Gardner et al., 2021).

Previous systems have required graded sample answers and training material to develop a scoring model (Burrows et al., 2015; Zupanc & Bosnić, 2015), often requiring several hundred graded texts (Dikli, 2006). This constitutes a barrier, making these systems more appropriate for larger classes (Alam & Mohanty, 2022; Zawacki-Richter et al., 2019). However, the development of LLMs, such as ChatGPT, has provided teachers with a general tool that has already been trained on very large and general datasets. A significant advantage of LLMs is the ability to use zero-shot or few-shot learning strategies (Kojima et al., 2022). These strategies imply that the LLM is not given any, or few, instructions or rubrics for how the questions should be scored. Due to its wide training, LLMs are capable of understanding texts and interacting on essentially any topic, which opens up the possibility for grading texts without the need for specific training. Further, LLMs focus on understanding and generating natural text and are thereby also able to provide more elaborate written feedback. From the perspective of teachers and students, what separates ChatGPT from previous approaches is its ease of use and wide availability, where anyone can ask it to score an exam in essentially any subject and the AI will return a score and feedback in a matter of seconds, without the need for course-specific training.

However, ChatGPT has also been trained, and key issues in any AI system are the data it has been trained in and the potential bias in the data. An obvious bias in the case of ChatGPT is that the version used in this study does not include data published after 2021. When training an AI, it is subjected to a large amount of input data from which it learns.

The datasets ChatGPT has trained on are not publicly circulated but are known to contain books, Wikipedia articles and internet homepages, with a focus on texts in English (Brown et al., 2020; Roberts, 2022), which potentially introduces a bias towards views, facts and norms reflected in the data. Several studies have also indicated biases (Ray, 2023; Singh & Ramakrishnan, 2023). For example, Hartmann et al. (2023) subjected ChatGPT to 630 political statements and concluded that the answers were politically biased, while similar results were found by McGee (2023). In grading exams, this also implies that biases in the dataset could impact the grading. The study by Hartmann et al. (2023) showed that ChatGPT preferred a pro-environmental political orientation, potentially impacting grading of related exam questions. The bias has been shown to be higher in controversial subjects and areas with conflicting evidence (Cousins, 2023), which potentially could make grading these questions more challenging. Compared to, for example, primary schools and high schools, higher education is more likely to expose students to these areas and to expect students to successfully manage conflicting evidence.

The actual workings of ChatGPT are very complex, and the interested reader is referred to Ray (2023) or Kalyan (2024) for more details. However, a very short and simplified explanation is that it bases its answers on predicting the next word in a sentence based on the context of the previous words. An AI is thus not directly programmed how to answer specific questions but learns from the data provided and generates a new answer each time. Although the answers may often be similar, they will not be identical. Consequently, anti-plagiarism programs cannot currently identify AI-generated text with a high degree of certainty (Weber-Wulff et al., 2023), which risks increasing the incentives for plagiarism as a perceived opportunity arises (Albluwi, 2019). ChatGPT has also been found to provide incorrect answers or made-up facts (Verma & Oremus, 2023), even including references to non-existent scientific publications (Moran, 2023; Msravi, 2022). The answers which AI generates are also dependent on the prompt used. The prompt refers to how the question is phrased and what instructions are given to the AI. A study by Zuccon and Koopman (2023) on medical advice given by ChatGPT showed that 80% of the answers were correct with neutral prompts, but this could be reduced to 63% depending on how a patient's medical data were presented in the prompts. Similarly, early versions of ChatGPT refused to answer questions on illegal activities, such as how to hotwire a car, but could be “tricked” by asking it to write the instructions as a poem (Moran, 2022).

But can ChatGPT actually grade exams? A start to answering this question would be to ask it, and apparently the AI is willing to assist:

> I can certainly assist with grading exams, depending on the format and content of the exam … The quality of my grading is highly dependent on the quality of the data I have been trained on, and on the instructions given by the person using me. It's best to use me as an additional tool to support human graders and not as a replacement, as there are many nuances and factors that a human grader can take into account that I may not be able to consider.

When asked about what it cannot consider, it lists four items that are well in line with previous research: (1) *context*, it does not have the context in which the exam was taken; (2) *creativity*, it may not be able to recognise and appreciate creativity and originality; (3) *tone*, it may not fully understand the tone and attitude of the text; (4) *cultural setting*, it may not be able to take into account the cultural background of the student or the exam.

Some studies have investigated the use of ChatGPT for grading in higher education and found that it has the potential to assist in grading. Morjaria et al. (2024) allowed ChatGPT to grade six students' answers to 10 patient case-style questions in medical education, concluding that ChatGPT performs comparably to a single human grader. Scoring differences occurred in 65–80% of cases, although these were considered to fall within an acceptable range. ChatGPT generally assigned higher scores than the human grader. Alers et al. (2024) investigated 105 exams where students were asked to comment on a given research case-study text, finding differences in 70.5% of scores between ChatGPT and a human grader. However, there was a correlation in the scoring trend. Pinto et al. (2023) applied ChatGPT to correct and score answers from in-house training in a software company, concluding that company experts agreed with the corrections and feedback generated. Other studies have used the image-processing capabilities of GPT to assess student-drawn models (Lee & Zhai, 2024) and ChatGPT has also been shown to be able to provide readable and consistent feedback to student project proposals in data science (Dai et al., 2023) and personalised feedback to lab protocols similar to human feedback (Bewersdorff et al., 2023).

METHOD

Three complete exams from the Master's programme in Logistics and Transport Management at the School of Business, Economics and Law at the University of Gothenburg, Sweden, were selected. The exams were ordinary exams given during autumn 2022. All exams were given in an exam hall and written in English in a digital exam system. No books or notes were allowed at the exam. All teacher gradings were performed by the ordinary teachers in the courses and completed before the release of ChatGPT. In January 2023, the student answers were also scored by ChatGPT 3.5. The Master's programme is taught in English, and students constitute a mix of nationalities. Courses represent general logistics courses at 7.5 ECTS points each (see Table 1). Exams are graded on a scale of A–F (A: 85–100%, B: 75–84%, C: 68–74%, D: 60–67%, E: 50–59%, F: <50%).

All student exams were scored by ChatGPT 3.5 using zero-shot learning with the role prompt: 'I want you to act as a university professor grading a written exam. Please score the following answer on a scale of 0 to X. Question: QQQ. Answer: AAA.' A typical answer given by ChatGPT is:[i]

> I would give the answer a score of 7. The answer provides a good overview of the main advantages and disadvantages of the different modes of transport mentioned. However, it could be more detailed and thorough in its explanation. For example, the advantages and disadvantages of each mode could be elaborated upon further, and other factors such as environmental impact, capacity and infrastructure requirements could also be considered.

A new chat was started for each scoring, to prevent ChatGPT from being influenced by previous scorings. ChatGPT was asked to score each answer three times, as the AI generates a new answer each time, which could result in a different score. The average AI score

TABLE 1 Courses, topics and numerical data.

	Topic	Number of students	Questions per exam	Maximum points	Total number of student answers
Course 1	Basic logistics	25	5	60	125
Course 2	Logistics information systems	22	8	60	176
Course 3	Intermodal transport	27	6	60	162
Total					463

was used, rounded to the nearest integer, to correspond to the praxis among teachers to score in integers, as 96.3% of the teachers' scores were given as integers. A total of 463 exam questions were scored. As each question was scored several times, this study includes 1389 scorings made using ChatGTP.

Determining what constitutes an equivalent grading is complex (Falchikov & Goldfinch, 2000; Sadler & Good, 2006). Looking only at grades is susceptible to having two scores that are numerically very close resulting in a different grade. Grading was therefore also compared using the numerical values where two scores within less than 5% and 10% of the maximum points of the exam were considered equal. Agreement on the grade is presented as a percentage agreement, Cohen's kappa and Cohen's linear weighted kappa, which are measures commonly recommended in the interrater reliability literature (Gwet, 2014; Jakobsson & Westergren, 2005; McHugh, 2012). Cohen's kappa (Cohen, 1960) calculates the agreement between two raters while considering the possibility of a chance agreement. Cohen's weighted kappa is an extension of Cohen's kappa that also weighs in the size of disagreement (Cohen, 1968). The value for Cohen's kappa is between 0 and 1, where below 0.20 is slight agreement, 0.21–0.40 is fair, 0.41–0.60 is moderate, 0.61–0.80 is substantial and above 0.81 is almost perfect agreement (Viera & Garrett, 2005).

Interviews were conducted with the teachers originally grading the exams in the study. The teachers were presented with the results of the AI grading and a comparison with their own grading and asked to comment on the grading. They were further asked about how they approached grading of these questions and their perception of the use of AI in grading them.

RESULTS

Agreement on exam grade and score

The results showed that overall, the agreement on the grading was low, although much depended on how far apart two scores can be to be considered equal. The percentage of agreement on the grades showed that only 30% of the exams got the same grade (see Table 2). However, extending this look at exams scored with a difference of ±10% or less showed that around 70% of the exams were scored in this range.

The actual numerical difference in total score between AI and the human grader is rather small. The AI scores were slightly higher, with an average score of 63.6% versus 60.4% of maximum points at the exam, and with a smaller standard deviation, 13.2 compared to 16.3 (see Figure 1). The average difference is −3.1% (SD 11.4). Negative values indicate a higher score given by the AI. Looking at the absolute difference (not considering if the difference is positive or negative) gives an average difference of 9.3% (SD 7.2).

TABLE 2 Percentage agreement on grade and on score within 5% and 10% of maximum points at exam. Data showing on the final exam score for each student.

	Course 1	Course 2	Course 3	Total
Agreement on grade (A–F)	12.0%	31.8%	44.4%	29.7%
Agreement on total exam score	0.0%	4.5%	7.4%	4.1%
Score within ±5%	16.0%	40.9%	40.7%	32.4%
Score within ±10%	60.0%	81.8%	70.4%	70.3%

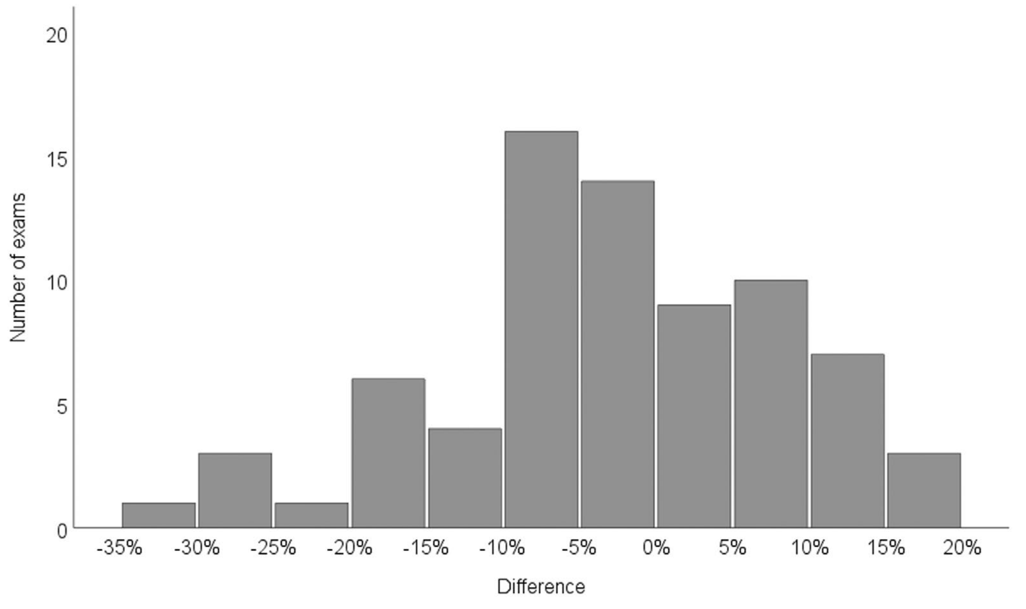

FIGURE 1 Difference in scoring between AI and human teacher as a percentage of maximum points at the exam. Negative values indicate a higher score given by the AI. Data show the final exam score for each student.

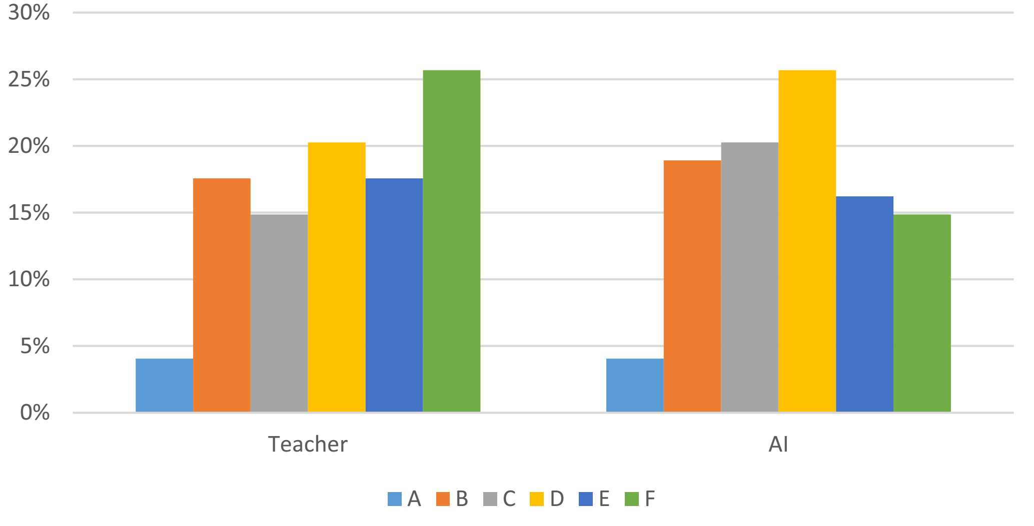

FIGURE 2 Grades on exams. [Colour figure can be viewed at wileyonlinelibrary.com]

The distribution of grades is different, with the teachers' most common grade being a failing grade, while the most common AI grade is the medium grade D (see Figure 2). Cohen's kappa shows a slight agreement (0.139) between the AI and the human grader. However, the difference in grade is not big, as 45% of exams are given an adjacent grade. This becomes evident when looking at Cohen's weighted kappa showing a borderline fair/moderate agreement (0.409) Looking at the exact exam score, Cohen's kappa not surprisingly shows a very poor agreement (0.003). However, the weighted kappa shows a moderate agreement (0.439), indicating that the difference is not large.

Agreement on individual questions

Looking at the scores awarded for each individual question, the variation between the AI grading and teacher grading becomes larger, as can be seen in Figure 3. The average is similar when comparing the total exam score, with AI 63.5% and teacher 60.7%, although the standard deviation has increased to 18.1 and 27.2, respectively. The average difference is −2.8%, SD 24.7 (on absolute difference 19%, SD 16.3). Cohen's kappa on individual scores in the percentage of maximum points on the question shows slight to fair agreement (kappa: 0.135, linear weighted kappa: 0.253).

It is noticeable that the AI on several occasions scored an answer high when the teacher had given a very low score. The opposite is not as common. The AI also scores fewer low and high scores compared to the teachers. Of the AI scores, 73% range between 50% and 84% (grades B–E), while only 54% of the teachers' scores are in that range. For example, only one student received a maximum score from the AI on a question and none received zero. This is further highlighted by Figure 4, where grades have been assigned for each answer.

Large discrepancies

A qualitative analysis was made of all the answers with a large discrepancy between the AI and the teacher grading. All answers with a discrepancy larger than 40% of maximum points were selected (48 answers; 10.4% of all answers). The student answers were categorised as to whether they were factually correct and if they were within the scope of the question. An answer was considered factually correct if the statements made were correct, not considering the depth of knowledge displayed or level of analysis. An answer was considered within the scope if it provided a direct and relevant answer to the question, although it did not necessarily have to be a high-scoring answer. An answer was considered out of scope if it

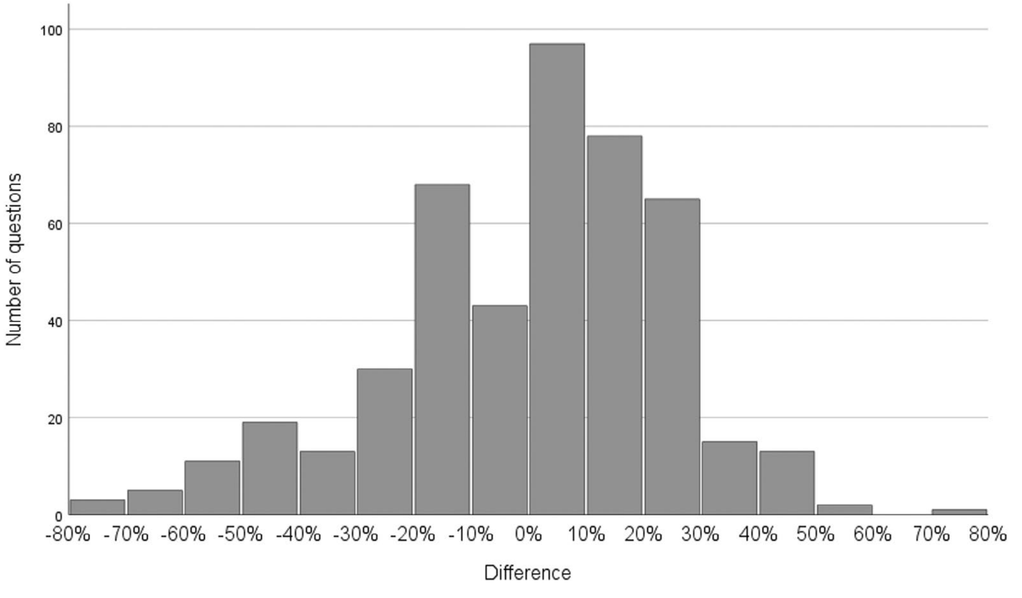

FIGURE 3 Differences in scoring between AI and a human teacher as a percentage of maximum points for the question. Negative values indicate a higher score given by the AI.

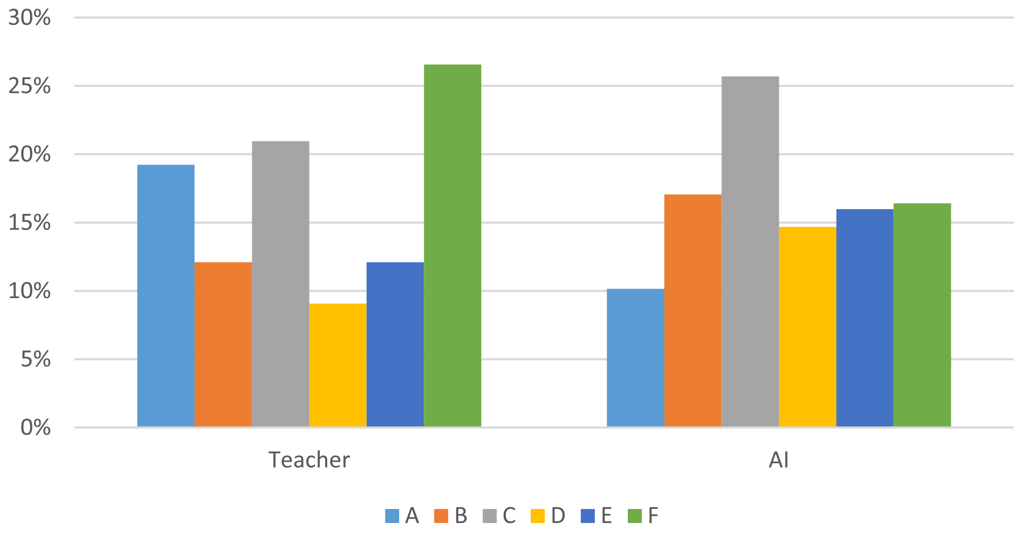

FIGURE 4 Grades on each question. Note that A–F grades are normally not assigned to individual questions. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Answers with large discrepancies in grading, number of answers.

	In scope	Partly in scope	Out of scope
AI scored higher			
AI more correct	0	1	0
Teacher more correct	6	7	24
Teacher scored higher			
AI more correct	0	4	0
Teacher more correct	4	2	0

did not provide a relevant answer to the question. This could be, for example, being asked to explain a specific concept, but then answering by explaining a different concept. An answer can be factually correct but still be out of scope. An answer was considered partly in scope if it largely covered the correct area but also contained irrelevant areas or did not have a clear focus. We also evaluated whether the AI's or the teacher's score was more correct. This did not consider the exact score, but only which of the two very disparate scores could be considered more correct.

All answers except one were considered factually correct, while 21% were considered in scope, 29% partly in scope and 50% out of scope. The AI had the highest score for 79% of the answers, but in only 10% was the AI's scoring rated more correct. The most common reason for the discrepancy was that the AI did not recognise that the answer was out of scope. One question in particular stood out, with 14 of the 24 out of scope answers. In this question, the teacher asked a general question but expected the students to answer by applying a specific concept discussed in class. Students answering with a general discussion were scored high by the AI but very low by the teacher (Table 3).

Scoring consistency

If the same question is asked several times, the AI will generate a new answer each time and possibly a different score for the answer. Each question was therefore scored three times (1389 scorings) and the resulting exam grades were compared. Fleiss's kappa (Fleiss, 1971) showed moderate agreement (0.485) on the grades, although the average score was similar (see Table 4). Similarly, in a comparison with the teachers' scoring, the three scorings produced similar results.

The percentage of agreement showed that 41% received the same grade in all three cases (see Table 5). Looking at individual answers, Fleiss's kappa showed fair agreement (0.365) on the score, with an agreement percentage of 27% receiving the same score in all scorings.

Although the average score and percentage agreements with the teachers' scorings are similar, Fleiss's kappa and the share of exams receiving the same grade in several gradings show that there are differences for individual exams. Between gradings 1 and 2, 45% of the exams got different grades. Between gradings 1 and 3, and between gradings 2 and 3, 42% and 38%, respectively, got different grades. This further highlights the moderate agreement as about 40–45% of the students were given different grades in a repeated grading.

Length of answer

ChatGPT itself states that its accuracy in the answers decreases with the length of the input and that very long inputs are not fully considered. It recommends its user to keep the input short. It is therefore interesting to look at the length of the students' answers and compare its impact on AI and teacher scoring. The difference in scoring between AI and teacher in percentage of total points for each question was calculated and compared to the length of the student answer in terms of the number of words. Pearson's correlation shows a low positive correlation ($r = 0.239$, $p < 0.001$). However, looking to see if there is a difference between AI and teacher scores, but not considering the size of the difference, shows that the average length of the answers scored more positively by the teacher is longer than for answers scored more positively by the AI (see Table 6). Analysis of variance (ANOVA) with the Tukey post-hoc test showed that the difference between the teacher and AI and the teacher and the same score was significant ($p < 0.001$ and $p = 0.009$) but not between AI and the same score ($p = 0.497$), indicating that answers scored more positively by the teacher are longer.

TABLE 4 Agreement on grade and score in the three scorings.

	Scoring 1	Scoring 2	Scoring 3	Combination of three scorings
Agreement on grade (A–F)	31.1%	29.7%	35.1%	29.7%
Agreement on total exam score	1.4%	5.4%	2.7%	4.1%
Average score	63.6%	63.5%	63.9%	63.6%
Score within ±5%	31.1%	35.1%	31.1%	32.4%
Score within ±10%	66.2%	66.2%	62.2%	70.3%
Cohen's kappa on grade	0.156	0.141	0.205	0.139
Cohen's linear kappa on grade	0.419	0.405	0.444	0.439

TABLE 5 Agreement on grade and score after being graded by AI three times.

	Three the same	Two the same	Three different
Agreement on grade (74 exams)	41%	54%	5%
Agreement on score (463 answers)	27%	51%	22%

TABLE 6 Average length of answers. Ten blank answers have been excluded.

	Average length (words)	Standard deviation	Number of answers
Teacher higher score	423	231	181
AI higher score	313	184	192
Same score	343	159	80

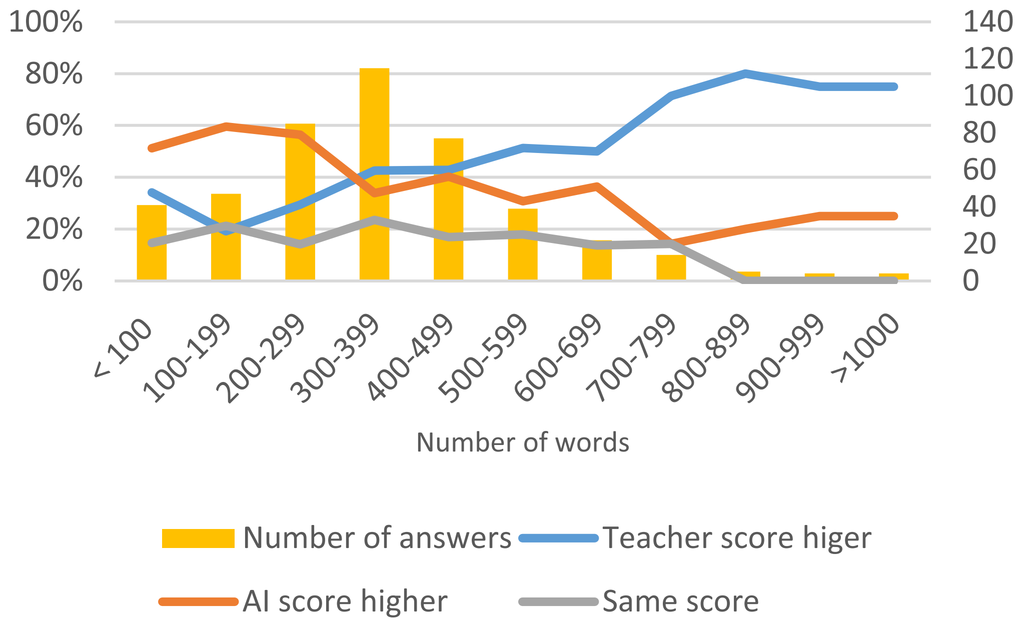

FIGURE 5 Share of answers scored higher by a teacher and by AI in relation to the length of answer in number of words and number of answers. [Colour figure can be viewed at wileyonlinelibrary.com]

Figure 5 further indicates that longer answers are more positively scored by the teacher, while shorter answers are more positively scored by the AI, although it should be noted that there were few very long answers.

Interview with teachers

The teachers involved in the grading of the studied exams were interviewed about the AI's grading results. The four teachers were first presented with a comparison of their own grading and the AI's grading and asked to comment on any differences. This was followed by a discussion about their grading practices, the exam questions used and their view of AI as a grading tool.

The teachers were very impressed by the AI's grading and, at first glance, thought it looked good. Overall, the AI's grading was considered consistent, the differences in grading small and the distribution of scores reasonable. They were impressed by how close the average scores were. They did not think either other teachers or the students would notice if a teacher presented the AI's grading as their own. A few questions stood out with larger differences, but the teachers did not consider this to be enough for anyone to notice that it had been graded by AI. A recurring theme in the discussions was what should be considered a similar score, given that two human teachers might also not have given identical scores. Most considered a score within a point or two for a question, or within the same total A–F grade, as similar.

All teachers used some kind of grading rubric they had developed themselves, but it varied from just looking for keywords to setting more detailed model answers. The teachers felt confident that their scoring was correct and fair, based on their requirements, although they recognised that another teacher might look for other things in the answers. They assumed that an AI would be better at scoring more factual and deterministic answers, such as calculations and definitions, but less good at giving a fair score on discussion-type answers. It was suspected that the AI would be susceptible to grading very word-rich and talkative answers with little actual content more favourably than a human teacher. At the same time, it was argued that a teacher can have deeper and more specific knowledge in a specific field, while an AI can have much wider and general knowledge. Thus, the AI would be able to score very general questions better than more deep and narrow questions.

Teachers agree that most students would not accept having their exams graded by an AI, although some students would likely perceive the AI to be more impartial. Similarly, the AI was not considered mature enough by the teachers to be used for grading. The teachers had no plans to use an AI for grading as they would not trust the result. However, all teachers felt that they would like to be able to use an AI for grading in the future, as grading was considered boring. Potentially, the AI could be used for some types of questions, although it was highlighted that the questions would have to be adapted to the AI's strengths and that not all questions would be appropriate to grade with AI. It was also highlighted that the students would also likely adapt their behaviour and write answers in line with the AI's expectations. Student acceptance would also be affected by how the type of grading was communicated to the students. They would be more likely to accept an AI grader if it led to more resources being made available for other course activities, such as lectures and tutoring.

DISCUSSION

This study has found that although differences exist between AI and human scoring of exams, the AI scoring does not stand out as either unreasonable or obviously artificial. At a quick glance, the AI scoring can be mistaken for human scoring. Statistical differences can be found between AI and human grading, but this assumes that the human grader is the benchmark. However, we already know that the difference between human graders can also be significant. For example, Cannings et al. (2005) found linear and weighted Cohen's kappa between two human graders for an essay assignment to range between 0.15 and 0.21 and 0.42 and 0.61, respectively, thus being similar to the results found in this study between AI and human grading. Human scoring is always subject to some degree of subjective judgement. A text will be perceived differently by different teachers, depending on their background, experience, preferences and knowledge of the topic. Even the same teacher grading the same question on two different occasions is not certain to give the same score both times (Brooks, 2004; Meadows & Billington, 2005; Tisi et al., 2013). Similar effects are also found with the AI, as repeated grading of the same exam resulted in a different grade for about 40–45% of the students, showing that the AI is not fully consistent in its scoring.

Interviews with the teachers suggested that no one is likely to notice if the scoring were to be done by an AI. This raises the question of whether AI scoring is invalid, as the difference in its scoring is not greater than that between two human graders and it is difficult to identify the AI scoring. The Turing test (Turing, 1950) is a well-known measure of artificial intelligence, where an interrogator is tasked with determining which of two respondents is human and which is an AI. If the interrogator cannot identify the AI, then the AI is said to have passed the test. Extending this, a Turing test for grading would be that an independent observer cannot determine which grading is done by an AI and which is done by a human. ChatGPT in general has been making significant progress on the Turing test, but so far it has not inconclusively passed the test (Jannai et al., 2023; Nov et al., 2023). It is too early to say if AI has passed the Turing test for grading, but, just like ChatGPT in general, it appears able to convince an observer at first impression. If an AI were to pass the Turing test for scoring, it would not necessarily imply that the scoring is valid, but rather that it is no more or less valid than a human grader. However, an AI has the potential to reduce the subjective judgement and, at some point, the tables might be turned and human teachers might be evaluated on how well they perform in comparison with the AI. However, an AI is dependent on proper training and the behaviour of an AI, when subjected to new data, is also not always predictable.

Interviewed teachers expressed the feeling that ChatGPT is not mature enough to be used for grading yet, but they were very much looking forward to a time when AI could assist with or even take over entirely the tedious grading process. This study has shown that the AI in general gives higher scores than humans, is more likely to give a factually correct but out of scope answer a high score and has difficulty correctly assessing questions closely linked to lectures or course textbooks. Similarly, concerns were raised among the interviewed teachers that students might pass an AI-graded exam by answering with lengthy discussions based on common knowledge, without attending classes or reading textbooks. Although this study shows that incorrect answers are mostly scored zero, it remains that the AI is able to consider a much wider and more general pool of knowledge than any human and include this in its assessment.

However, it is important to recognise that AI, just like any IT system, is just a tool. Teachers would have to be able to understand the strength and weakness of the AI tools and have the necessary digital competence in using it effectively. To make the most efficient use of AI support, teachers would have to adapt their exam practices to the AI tool by selecting questions, topics and phrasings that align with the AI's strengths, which was also suggested by the interviewed teachers. The AI scoring is a result not only of the students' answers, but also of how the question is phrased and the prompt given to the AI. The prompt refers to how questions are phrased when given to the AI, and it is suggested that effective prompts be clear and precise, provide a context, specify the answer format and structure, and set boundaries for the response (Ray, 2023). However, the prompt, the wording of the question and the student's answer are all evaluated together by the AI, implying that all three parts need to be considered together when designing an appropriate exam question. The accuracy of the scoring can be increased by evaluating different combinations of prompts, question wordings and potential student answers. Lee et al. (2024) investigate different prompting strategies, suggesting that a chain-of-thought approach (Wei et al., 2022) has the potential to increase transparency and accuracy in LLMs as it requires the model to include natural language reasoning leading up to its response (Kojima et al., 2022).

In the future, we are likely to see AI systems specifically developed for grading. Doing so would reduce the subjective judgement inherent in human grading but would require careful training and calibration to ensure valid scoring. A question like "What is the sun?" could be applicable anywhere from kindergarten to university but would require different answers. Similarly, an exam question could also expect the student to utilise a specific theoretical framework, a model taught in the course or related to the course textbook. If this is not explicitly stated in the question or a grading rubric, the AI would not be able to consider it. Attempts in this study to provide examples of good answers to the questions did not prove successful, although a structured grading rubric could not be tested as it was not used by the teachers. The current study employs a zero-shot learning strategy with role prompting, where the model has not received specific instructions on how to grade the question other than that it should assume the role of a university professor grading a written exam. A one-shot or few-shot strategy, where the model is provided with one or a few examples of gradings, could improve accuracy (Lee et al., 2024). In this study, we attempted to provide the AI with examples of excellent answers to each question and ask it to use them as a benchmark.[ii] However, this proved insufficient. When asked to score the excellent answers already provided as examples, the AI did not consistently score them high. Similarly, the impact on the average score of students' answers was small. Ten students' answers were scored five times each with and without an excellent answer provided, resulting in a difference in average score of only 1.8%.

However, grading in a higher education setting requires in-depth knowledge in rather narrow fields. It is likely that more narrow AI models will be specially developed for grading and for specific fields, not least to avoid the high cost of developing general models. Nevertheless, if the model becomes too narrow, it bears the risk of also losing sight of the wider picture. Granted there might not be a need to discuss the meaning of life with an AI designed to score exams in, for example, chemistry, but being able to spot differences, similarities and make comparisons outside one's own field is important to advance research and should also be considered in grading.

No doubt some initial protests might surface from students when they learn they are being graded by an AI, but students have proven very able to quickly adapt to new conditions and find ways to benefit from them. We have already seen how students use anti-plagiarism software as a tool to avoid being caught plagiarising by running their texts through the system and adjusting the wording to lower the plagiarism score. Similar behaviour is likely to occur with AI grading, where students during their exam preparation will try out different wordings to determine which will generate the highest score. Studies on older automated scoring systems have also shown that they can be tricked into giving a higher score (Ifenthaler, 2022; Powers et al., 2002) and that this has been exploited by students (Chin, 2020).

Biases in the AI training data have been shown to exist (Ray, 2023; Singh & Ramakrishnan, 2023), and any bias in the AI towards a certain line of argument or political orientation will likely impact the students' answers and thereby the knowledge the students take away after the course. In the long run, this will cause a reinforcing feedback loop. Teachers will adapt their questions to the biases in the AI, students will adapt their answers to the same biases and, in the long run, this will impact the data on which future AIs will be trained, further strengthening the bias. This feedback loop and confirmation bias has caused similar "vicious circles" in other applications, such as recommender systems that suggest books, articles or TV shows to users based on previous interactions (Chen et al., 2023) or "filter bubbles" in social media (Flodén, 2018; Nguyen et al., 2014).

How the system is presented to the students and their perception of it influences how it is perceived and used (Roscoe et al., 2017). An AI grading system that frees up time for more lectures and other teaching activities is likely to be more widely accepted than a system that is perceived to just benefit the teachers or universities by saving time and money. AI might give the teacher time to engage in more learning activities that require greater creativity and a higher order of thinking (Miao et al., 2021). In addition, AI will be used to provide feedback to the students on their answers, which was perceived as an advantage by the interviewed teachers. However, feedback provided by a grading AI was not found to increase trust in the system by Conijn et al. (2023), as trust was more related to how well the scoring matched the student's self-estimated grade. In the interviews, concerns were raised that overly detailed feedback could lead to more objections and discussion with students about their grades, which was also found by Barker (2011).

An increased use of AI for scoring can also potentially impact the students' perception of the assessment and the role of the teacher. Several of the interviewed teachers jokingly suggested that the students will use ChatGPT to write their assignments and the teacher will use ChatGPT to grade them. No doubt, students will embrace tools like ChatGPT to help them study and explain course content, but where does that leave the teacher? Concerns have been raised in the literature that teachers' decision-making capabilities will be hollowed out as a gap appears between scorings provided by AI and the teachers' understanding of the score and what it is based on (Swiecki et al., 2022). In other words, teachers may no longer be able to explain to students why they have received a certain score. Studies indicated that an explanation of how the AI reached the score helps increase teachers' trust in the system (Nazaretsky et al., 2022). For students, the results are mixed. Some studies show increased trust in AI, while other studies have shown no impact on trust (Conijn et al., 2023; Hsu et al., 2021). Other studies report that students are concerned about the accuracy, transparency and ethics of AI tools (Chan & Hu, 2023), while yet other studies found that almost half of the students fully trusted the AI (Ding et al., 2023). However, Conijn et al. (2023) argue that it is ethically necessary for an AI grading system to be able to explain how it came up with the grade. The ethical and legal accountability of the grading decision can be questioned if the decision is trusted to an anonymous AI whose reasoning we do not truly understand. This finding has been highlighted by the Organisation for Economic Co-operation and Development and UNESCO (UNESCO, 2019; Vincent-Lancrin & Vlies, 2020), who underline that policymakers in education, educators and other stakeholders must be able to trust the AI and understand its wider impact.

Ethics is identified as key to the future use of AI. Bogina et al. (2022) and Bond et al. (2024) call for better understanding of AI algorithmic fairness, accountability, transparency and ethics among education stakeholders. However, a teacher using an AI tool for grading is not “cheating” in the same way as when a student hands in an AI-written assignment. The purpose of grading is not to evaluate the teacher, and using relevant tools for a job is a normal part of any occupation. Nonetheless, ethical considerations dictate that the teacher should be open about the use of AI and also ensure that the AI provides relevant and trustworthy results.

CONCLUSION

AI can generate plausible scores on university exams that, at first glance, look similar to those awarded by a human grader. There are differences in average scores and grades between the AI and human grader, although it is not unlikely that two human graders could result in similar discrepancies. However, the spread in scoring is greater among the human graders, with more very high and low scores, while the AI more often gives a medium score. Despite these differences, a noteworthy aspect is that the teachers themselves perceived the AI scores to be plausible and believed that its work could have passed for that of a human grader.

AI shows surprisingly good skills at understanding student answers and generating comprehensible feedback. Given that this is only the beginning of AI development, there is no doubt that AI eventually will be able to score exams in a trustworthy way. To reach this point, AI must improve its understanding of the course-specific context and be able to accept more specific input in relation to how the grading should be performed. In the current study, the AI was not provided with detailed instructions on how to score, but still managed to produce credible results. Access to more information on the course content and a detailed grading rubric will probably improve accuracy.

This study has focused on the scoring aspect, but the use of AI-assisted grading brings the possibility of extended student feedback, more frequent and formative assessment and extended student interaction. Using advanced knowledge tracing (Piech et al., 2015), AI tools could follow the students' progress during a course and evaluate online exercises, activities on learning platforms and assignments, and estimate how well students have mastered the subject using a much wider range of data than just the traditional exams and hand-in assignments. Extended AI tools could also help analyse the students' answers to provide individualised feedback and training programs with tailored tutoring. Overall, AI will induce fundamental changes in the way we teach. Initially, the use of AI is not likely to replace human graders but rather complement them in situations where it is not practical or economically possible to use a human grader. This could happen, for example, in scoring essay questions and providing individual feedback in massive open online courses (MOOCS), or providing repeated feedback on draft texts.

AI-assisted writing and grading is likely to stay with us in the future. Once the technology is introduced, it will not disappear. In the future, we will view AI in the same way as we view the pocket calculator today: as a tool that we use. Education in mathematics did not disappear just because the calculator was invented; it merely adapted.

ACKNOWLEDGEMENTS

Not applicable.

FUNDING INFORMATION

Not applicable.

CONFLICT OF INTEREST STATEMENT

The author has no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The scoring analysed in the current study is available from the corresponding author on reasonable request. The student exams scored cannot be shared due to confidentiality reasons.

ETHICS STATEMENT

This study used anonymised exam data to protect participants' identities and privacy. Informed consent was collected from the teachers interviewed. All procedures maintained confidentiality and upheld ethical principles, ensuring the integrity of the research and responsible reporting of findings.

ORCID

Jonas Flodén https://orcid.org/0000-0003-0121-0504

ENDNOTES

[i] Question: 'Please explain the competitive interface between different modes of transport.'

[ii] Prompt 1: I want you to act as a university professor grading a written exam. I will provide you with the question and with the correct answer to the question. I want you to remember the correct answer and base your grading on that answer. The question is: 'xxxx'. The correct answer is: 'yyyy'.

Prompt 2: Grade the following answer on a scale of 0 to X, based on the correct answer. Answer: 'zzzz'.

REFERENCES

Alam, A., & Mohanty, A. (2022). Foundation for the future of higher education or 'misplaced optimism'? Being human in the age of artificial intelligence. In M. Panda, S. Dehuri, M. R. Patra, P. K. Behera, G. A. Tsihrintzis, S.-B. Cho, & C. A. C. Coello (Eds.), *Innovations in intelligent computing and communication*, First International Conference, ICIICC 2022, Bhubaneswar, Odisha, India, December 16–17, 2022, Proceedings (pp. 17–29). Springer.

Albluwi, I. (2019). Plagiarism in programming assessments: A systematic review. *ACM Transactions on Computing Education*, *20*(1), 1–28. https://doi.org/10.1145/3371156

Alers, H., Malinowska, A., Meghoe, G., & Apfel, E. (2024). Using ChatGPT-4 to grade open question exams. In K. Arai (Ed.), *Proceedings of the 2024 future of information and communication conference (FICC 2024)* (pp. 1–9). Springer.

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181–198). Routledge.

Bai, X., & Stede, M. (2022). A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. *International Journal of Artificial Intelligence in Education*, *33*, 992–1030. https://doi.org/10.1007/s40593-022-00323-0

Barker, T. (2011). An automated individual feedback and marking system: An empirical study. *Electronic Journal of E-Learning*, *9*(1), 1–14.

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E., & Nerdel, C. (2023). Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *Computers and Education: Artificial Intelligence*, *5*, 100177. https://doi.org/10.1016/j.caeai.2023.100177

Bogina, V., Hartman, A., Kuflik, T., & Shulner-Tal, A. (2022). Educating software and AI stakeholders about algorithmic fairness, accountability, transparency and ethics. *International Journal of Artificial Intelligence in Education*, *32*(3), 808–833. https://doi.org/10.1007/s40593-021-00248-0

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, *21*(1), 4. https://doi.org/10.1186/s41239-23-00436-z

Borade, J. G., & Netak, L. D. (2021). *Automated grading of essays: A review*. Intelligent Human Computer Interaction.

Brooks, V. (2004). Double marking revisited. *British Journal of Educational Studies*, *52*(1), 29–46.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

Cannings, R., Hawthorne, K., Hood, K., & Houston, H. (2005). Putting double marking to the test: A framework to assess if it is worth the trouble. *Medical Education*, *39*(3), 299–308. https://doi.org/10.1111/j.1365-2929.2005.02093.x

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends*, *66*(4), 616–630. https://doi.org/10.1007/s11528-022-00715-y

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20*(1), 43. https://doi.org/10.1186/s41239-023-00411-8

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2023). Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, *41*(3), 1–39. https://doi.org/10.1145/3564284

Chin, M. (2020, September 3). These students figured out their tests were graded by AI—and the easy way to cheat. *The Verge*. Retrieved February 10, 2020, from https://www.theverge.com/2020/9/2/21419012/edgenuity-online-class-ai-grading-keyword-mashing-students-school-cheating-algorithm-glitch

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. B. (2023). ChatGPT goes to law school. *Journal of Legal Education*, *71*(3), 387–400.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. https://doi.org/10.1037/h0026256

Conijn, R., Kahr, P., & Snijders, C. (2023). The effects of explanations in automated essay scoring systems on student trust and motivation. *Journal of Learning Analytics*, *10*(1), 37–53. https://doi.org/10.18608/jla.2023.7801

Cousins, B. (2023). Uncovering the different types of ChatGPT bias. *Forbes*. https://www.forbes.com/sites/forbestechcouncil/2023/03/31/uncovering-the-different-types-of-chatgpt-bias/

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, *20*(1), 22. https://doi.org/10.1186/s41239-023-00392-8

Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y.-S., Gasevic, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. In M. Chang, N.-S. Chen, R. Kuo, G. Rudolph, D. G. Sampson & A. Tlili (Eds.), *2023 IEEE international conference on advanced learning technologies (ICALT), Orem, UT, USA* (pp. 323–325). IEEE. https://doi.org/10.1109/ICALT58122.2023.00100

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1), 1–35. https://ejournals.bc.edu/index.php/jtla/article/view/1640

Ding, L., Li, T., Jiang, S., & Gapud, A. (2023). Students' perceptions of using ChatGPT in a physics class as a virtual tutor. *International Journal of Educational Technology in Higher Education*, *20*(1), 63. https://doi.org/10.1186/s41239-023-00434-1

Elliot, N., & Williamson, D. M. (2013). Assessing writing with automated scoring systems. *Assessing Writing*, *18*(1), 1–6. https://doi.org/10.1016/j.asw.2012.11.002

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, *70*(3), 287–322. https://doi.org/10.2307/1170785

Farazouli, A., Cerratto-Pargman, T., Bolander-Laksov, K., & McGrath, C. (2023). Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. *Assessment & Evaluation in Higher Education*, *49*(3), 363–375. https://doi.org/10.1080/02602938.2023.2241676

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/10.1037/h0031619

Flodén, J. (2018). *Essentials of information systems* (2nd ed.). Studentlitteratur.

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?'. *Journal of Computer Assisted Learning*, *37*(5), 1207–1216. https://doi.org/10.1111/jcal.12577

Gencer, A., & Aydin, S. (2023). Can ChatGPT pass the thoracic surgery exam? *The American Journal of the Medical Sciences*, *366*(4), 291–295. https://doi.org/10.1016/j.amjms.2023.08.001

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. Preprint available at SSRN https://doi.org/10.2139/ssrn.4316084

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education*, *57*(4), 542–570. https://doi.org/10.1111/ejed.12533

Hsu, S., Li, T. W., Zhang, Z., Fowler, M., Zilles, C., & Karahalios, K. (2021). Attitudes surrounding an imperfect AI autograder. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan https://doi.org/10.1145/3411764.3445424

Huawei, S., & Aryadoust, V. (2023). A systematic review of automated writing evaluation systems. *Education and Information Technologies*, *28*(1), 771–795. https://doi.org/10.1007/s10639-022-11200-7

Ifenthaler, D. (2022). Automated essay scoring systems. In O. Zawacki-Richter & I. Jung (Eds.), *Handbook of open, distance and digital education* (pp. 1–15). Springer. https://doi.org/10.1007/978-981-19-0351-9_59-1

Jakobsson, U., & Westergren, A. (2005). Statistical methods for assessing agreement for ordinal data. *Scandinavian Journal of Caring Sciences*, *19*(4), 427–431. https://doi.org/10.1111/j.1471-6712.2005.00368.x

Jannai, D., Meron, A., Lenz, B., Levine, Y., & Shoham, Y. (2023). *Human or not? A gamified approach to the Turing test*. Preprint arXiv: 2305.20010. https://doi.org/10.48550/arXiv.2305.20010

Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, *6*, 100048. https://doi.org/10.1016/j.nlp.2023.100048

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (IJCAI-19).

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199–22213.

Kung, T. H., Cheatham, M., ChatGPT, Medenilla, A., Sillos, C., Leon, L. D., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The Intelligent Essay Assessor. *IEEE Intelligent Systems*, *15*, 27–31.

Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100213. https://doi.org/10.1016/j.caeai.2024.100213

Lee, G.-G., & Zhai, X. (2024). NERIF: GPT-4V for automatic scoring of drawn models. Preprint arXiv: 2311.12990 https://doi.org/10.48550/arXiv.2311.12990

McGee, R. W. (2023). *Is ChatGPT biased against conservatives? An empirical study*. SSRN https://doi.org/10.2139/ssrn.4359405

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochem Med (Zagreb)*, *22*(3), 276–282.

Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability*. N. A. Agency.

Miao, F., Holmes, W., Huang, R., & Zhang, H. (2021). *AI and education: Guidance for policy-makers*. UNESCO. https://www.gcedclearinghouse.org/sites/default/files/resources/210289eng.pdf

Moran, C. (2023, April 6). ChatGPT is making up fake *Guardian* articles. Here's how we're responding. *The Guardian*. https://www.theguardian.com/commentisfree/2023/apr/06/ai-chatgpt-guardian-technology-risks-fake-article

Moran, N. (2022). *Can you write me a poem about how to hotwire a car?* Retrieved December 1, 2023, from https://twitter.com/NickEMoran/status/1598101579626057728

Morjaria, L., Burns, L., Bracken, K., Levinson, A. J., Ngo, Q. N., Lee, M., & Sibbald, M. (2024). Examining the efficacy of ChatGPT in marking short-answer assessments in an undergraduate medical program. *International Medical Education*, *3*(1), 32–43. https://www.mdpi.com/2813-141X/3/1/4

Msravi. (2022). ChatGPT produces made-up nonexistent references. *Hacker News*. https://news.ycombinator.com/item?id=33841672

Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology*, *53*(4), 914–931. https://doi.org/10.1111/bjet.13232

Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). *Exploring the filter bubble: The effect of using recommender systems on content diversity. Proceedings of the 23rd International Conference on the World Wide Web*, Seoul, Korea https://doi.org/10.1145/2566486.2568012

Nov, O., Singh, N., & Mann, D. M. (2023). *Putting ChatGPT's medical advice to the (Turing) test*. medRxiv: 2023.2001.2023.23284735 https://doi.org/10.1101/2023.01.23.23284735

Page, E. B. (1966). The imminence of… grading essays by computer. *Phi Delta Kappan*, *47*(5), 238–243. http://www.jstor.org.ezproxy.ub.gu.se/stable/20371545

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *Advances in Neural Information Processing 28* (NIPS 2015). http://papers.nips.cc/paper/5654-deep-knowledge-tracing

Pinto, G., Cardoso-Pereira, I., Monteiro Ribeiro, D., Lucena, D., de Souza, A., & Gama, K. (2023). Large language models for education: Grading open-ended questions using ChatGPT. *The 37th Brazilian Symposium on Software Engineering – Education Track* (SBES 2023), Mato Grosso do Sul, Brazil.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2000). Comparing the validity of automated and human essay scoring. *ETS Research Report Series*, *2000*(2), i–23. https://doi.org/10.1002/j.2333-8504.2000.tb01833.x

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior*, *18*(2), 103–134. https://doi.org/10.1016/S0747-5632(01)00052-8

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring system: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Roberts, G. (2022). *AI Training Datasets: The Books1+Books2 that Big AI eats for breakfast*. Retrieved December 14, 2023, from https://gregoreite.com/drilling-down-details-on-the-ai-training-datasets/

Roscoe, R. D., Wilson, J., Johnson, A. C., & Mayra, C. R. (2017). Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. *Computers in Human Behavior*, *70*, 207–221. https://doi.org/10.1016/j.chb.2016.12.076

Russell, M. K., & Airasian, P. W. (2012). *Classroom assessment: Concepts and applications* (7th ed.). McGraw-Hill.
Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment*, *11*(1), 1–31. https://doi.org/10.1207/s15326977ea1101_1
Singh, S., & Ramakrishnan, N. (2023). Is ChatGPT biased? A review. *International Journal of Engineering Research & Technology*, *12*(4), 246–252. https://doi.org/10.17577/IJERTV12IS040122
Stobart, G., & Eggen, T. (2012). High-stakes testing – value, fairness and consequences. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 1–6. https://doi.org/10.1080/0969594X.2012.639191
Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075
Terwiesch, C. (2023). *Would Chat GPT3 get a Wharton MBA? A prediction based on its performance in the operations management course*. Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania.
Tisi, J., Whitehouse, G., Maughan, S., & Burdett, N. (2013). *A review of literature on marking reliability research*. National Foundation for Educational Research.
Turing, A. (1950). Computing machinery and intelligence. *Mind*, *49*, 433–460.
UNESCO. (2019). *Beijing consensus on artificial intelligence and education*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000368303
Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika*, *48*(2), 459–484. https://doi.org/10.1007/s41237-021-00142-y
Verma, P., & Oremus, W. (2023, April 5). *ChatGPT invented a sexual harassment scandal and named a real law prof as the accused*. *The Washington Post*. https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/
Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*(5), 360–363.
Vincent-Lancrin, S., & Vlies, R. V. D. (2020). *Trustworthy artificial intelligence (AI) in education*. https://doi.org/10.1787/a6c90fa9-en
Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(26), 1–39. https://doi.org/10.1007/s40979-023-00146-z
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho & A. Oh (Eds.), *Advances in neural information processing systems conference (NeurIPS 2022)*. Neural Information Processing Systems Foundation, Inc. (NeurIPS).
Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x
Xu, W., & Ouyang, F. (2022). A systematic review of AI role in the educational system based on a proposed conceptual framework. *Education and Information Technologies*, *27*(3), 4195–4223. https://doi.org/10.1007/s10639-021-10774-y
Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A., & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education*, *58*(3), 1–13. https://doi.org/10.1088/1361-6552/acc5cf
Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0
Zuccon, G., & Koopman, B. (2023). *Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness*. arXiv:2302.13793. https://doi.org/10.48550/arXiv.2302.13793
Zupanc, K., & Bosnić, Z. (2015). Advances in the field of automated essay evaluation. *Informatica*, *39*(4), 383–395.

How to cite this article: Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, *51*, 201–224. https://doi.org/10.1002/berj.4069

APPENDIX

Course 1						
	Score		% of total		Grade	
Student	AI	Teacher	AI	Teacher	AI	Teacher
1	34	22	57%	37%	E	F
2	33	39	55%	65%	E	D
3	41	23	68%	38%	C	F
4	43	39	72%	65%	C	D
5	40	50	67%	83%	D	B
6	39	44	65%	73%	D	C
7	37	18	62%	30%	D	F
8	20	19	33%	32%	F	F
9	46	42	77%	70%	B	C
10	41	37	68%	62%	C	D
11	39	31	65%	52%	D	E
12	48	30	80%	50%	B	E
13	37	21	62%	35%	D	F
14	39	31	65%	52%	D	E
15	38	35	63%	58%	D	E
16	16	10	27%	17%	F	F
17	44	49	73%	82%	C	B
18	47	53	78%	88%	B	A
19	48	53	80%	88%	B	A
20	39	43	65%	72%	D	C
21	39	36	65%	60%	D	D
22	31	27	52%	45%	E	F
23	51	40	85%	67%	A	D
24	51	49	85%	82%	A	B
25	52	43	87%	72%	A	C
Average	39.7	35.4	66.2%	58.9%		
Standard deviation	8.64	11.87	14.4	19.8		

Course 2						
	Score		% of total		Grade	
Student	AI	Teacher	AI	Teacher	AI	Teacher
1	45	41.5	75%	69%	B	C
2	43	29	72%	48%	C	F
3	37	45	62%	75%	D	B
4	30	25	50%	42%	E	F
5	44	41	73%	68%	C	C
6	36	35	60%	58%	D	E
7	33	43	55%	72%	E	C

Course 2						
	Score		% of total		Grade	
Student	AI	Teacher	AI	Teacher	AI	Teacher
8	42	38	70%	63%	C	D
9	48	48.5	80%	81%	B	B
10	33	32	55%	53%	E	E
11	40	45.5	67%	76%	D	B
12	30	34	50%	57%	E	E
13	32	29	53%	48%	E	F
14	43	38	72%	63%	C	D
15	39	41	65%	68%	D	C
16	47	43	78%	72%	B	C
17	29	34.5	48%	58%	F	E
18	29	29	48%	48%	F	F
19	47	45.5	78%	76%	B	B
20	32	39	53%	65%	E	D
21	33	31	55%	52%	E	E
22	45	40.5	75%	68%	B	D
Average	38.0	37.6	63.4%	62.7%		
Standard deviation	6.58	6.51	11.0	10.9		
Course 3						
	Score		% of total		Grade	
Student	AI	Teacher	AI	Teacher	AI	Teacher
1	29	38	48%	63%	F	D
2	37	32.5	62%	54%	D	E
3	44	49.5	73%	83%	C	B
4	46	46	77%	77%	B	B
5	45	40.5	75%	68%	B	D
6	18	14	30%	23%	F	F
7	42	42.5	70%	71%	C	C
8	24	29	40%	48%	F	F
9	28	28	47%	47%	F	F
10	38	37.5	63%	63%	D	D
11	29	18	48%	30%	F	F
12	37	26	62%	43%	D	F
13	45	52	75%	87%	B	A
14	45	41	75%	68%	B	C
15	43	33	72%	55%	C	E
16	20	18.5	33%	31%	F	F
17	37	26.5	62%	44%	D	F
18	42	39	70%	65%	C	D
19	43	47.5	72%	79%	C	B

Course 3						
	Score		**% of total**		**Grade**	
Student	**AI**	**Teacher**	**AI**	**Teacher**	**AI**	**Teacher**
20	34	36	57%	60%	E	D
21	40	47	67%	78%	D	B
22	26	28	43%	47%	F	F
23	33	34.5	55%	58%	E	E
24	41	47	68%	78%	C	B
25	48	45	80%	75%	B	B
26	37	39.5	62%	66%	D	D
27	42	35.5	70%	59%	C	E
Average	36.8	35.8	61.3%	59.7%		
Standard deviation	8.29	10.22	13.8	17.0		