Challenges and approaches related to AI-driven grading of open exam questions in higher education: Human in the loop

Author: Henry David Kurzhals
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT,
this research investigates the opinion of students regarding the use of Artificial Intelligence-based tools in the grading process of open exam questions in higher education. To close the gap in literature, this study considers the opinions of students, which have not been inspected in-depth before. A qualitative survey was conducted to identify which aspects/steps of the grading process students feel comfortable to be graded by Artificial Intelligence and which parts teachers should occupy. The results show, that most students can imagine AI to grade multiple-choice questions on its own as well as open questions under the supervision of teachers, while a minority of students do not accept AI in grading exams in higher education at all. These results imply that education and communication regarding AI-based tools for grading open exam questions in higher education need to be expanded and improved in the future in order to reach a common acceptance of it among students.

Graduation Committee members:
Dr. Daniel Braun
Dr. Patrizia Rogetzer

Keywords
Artificial Intelligence, machine learning, higher education, open exam questions

1. INTRODUCTION

Artificial Intelligence, in the following referred to as AI, is being described as the ability of computers to perform cognitive tasks, usually associated with human minds, particularly in the learning and problem-solving (Baker & Smith, 2019). AI has grown in popularity in a myriad of different industries in recent years. Among these are medicine and healthcare, transportation and logistics, robotics, the Internet of Things (IoT), Industry 4.0, finance, advertising, and especially data analysis. The number of companies (especially Startups) offering services or products based on or working with AI has grown immensely in recent years.

One of the sectors which also came in touch with the opportunities and challenges of AI, is the education sector, especially the higher education sector. AI is currently used to improve the learning process for students, as a feedback creator, or to support the teachers in their tasks. Popular AI-based applications on the market are “Copyleaks AI grader” or “Gradescope” (Brennen, 2020). Another tool called “Proctorio” is used to test the integrity of students, and automatic ID verification, and provides an admin dashboard and aggregates exam data as well as content protection and copy/download protection (Ahmad, Maabreh, Ghaly, Khan, & Qadir, S. 2022). Furthermore, AI-based automated scoring systems can support teachers in the assessment process (Kersting, Sherin, & Stigler, 2014). They do so by reducing teachers´ workload and helping them to focus their attention on critical issues such as timely intervention and assessment (Vij, Tayal, & Jain, 2020).

Although AI is capable of grading multiple-choice questions, the grading of open exam questions still has room for improvement before being implemented in higher education. That means these tools are currently in the progression stage. One of these tools is “EasyGrader”, a tool based on AI which focuses on grading open questions from exams in higher education. Tools like this need a high amount of training data from past exams to learn to create a proper assessment for the questions. Algorithms are used to focus on words and sentences regarding a certain topic. If these words or sentences appear in the answer to the open questions, the algorithm identifies them and accordingly grades the open question.

The principle behind it is machine learning, which can be described as the action of automatic improvement of computers through experience, lying at the intersection of computer science and statistics and the core of the AI and data science (Jordan & Mitchell, 2015).

A crucial point to take into consideration when speaking of the implementation of AI-based grading tools is the perception of students. Fairness and transparency need to be present. It is important to know what the students think of AI-based grading of open questions and where exactly in the grading process they can imagine AI doing the work and where they prefer leaving it to the teachers. The opinion of students is crucial since they are the subjects affected by the possible implementation of AI-based grading in the future.

Further, acceptance to change is better reached once trust is established, which can be done by being transparent and fair in designing, developing, and implementing the tool. Also, it can be done by educating and communicating to the students about AI-based grading tools, especially in the given situation where there might be a lack of information and knowledge. Once students are persuaded and resistance is overcome, one possible outcome might be that they will even help implement the tool in the future (Kotter & Schlesinger, 1989). AI-based grading tools can only reach the best version of themselves once this aspect is clarified. The displayed facts and analyses lead to the urge of examining students' knowledge and acceptance of AI-based grading of open exam questions in higher education and thus to the following research question for this paper:

2. RESEARCH QUESTION

“What is the opinion of students regarding which steps the AI-based grading tool “EasyGrader” in higher education should occupy in grading open exam questions and which step should be done by teachers?”

3. ACADEMIC RELEVANCE

The research problem of this paper is: “What is the opinion of students regarding which steps the AI-based grading tool “EasyGrader” in higher education should occupy in grading open exam questions and which step should be done by teachers?”. By finding an answer to this question, increased understanding and consequently acceptance of students regarding Artificial Intelligence might be created and spread.

Recently, many types of research aimed at analyzing AI´s role in education in general (Chen, Chen, & Lin, 2020). Also, Baker & Smith (2019) approached educational AI tools in a broad view. One of the few studies addressing Student´s opinions of AI-based grading tools in higher education was conducted by Sanchez-Prieto, Cruz-Benito, Theron Sanchez & Garcia-Peñalvo (2020). Although the perception of students has been initially reviewed, there is only little known of students´ opinions on which parts of grading in higher education should be done by AI and which parts by humans. Especially fairness and transparency are important factors to consider, which is why this research might contribute to that topic in literature and science. Furthermore, this paper might also serve as a basis for future research on AI in higher education.

4. PRACTICAL RELEVANCE

By publishing the research results, the information found could be of added value for scientists and developers of AI-based grading tools for open exam questions in higher education. Furthermore, the results can potentially facilitate the decision-making process of managers or policymakers (Toffel, 2016) in the implementation of Artificial Intelligence-based assessment tools in higher education in general. In the best possible outcome, the potential practical relevance of this research goes via a perceived research relevance by practitioners and the use of the research in practice directly to a realized research impact (Moeini, Rahrovani, & Chan, 2019).

Especially to the developers of the AI-based grading tool for open exam questions “EasyGrader”, this research might be of practical relevance in the further steps of developing and designing the tool and finally in the planning of implementing it into higher education assessment processes.

5. LITERATURE REVIEW

The literature existing regarding AI-based grading in higher education mainly focuses on the way AI supports teachers to work time-efficient or facilitates the grading process (Vij, Tayal, & Jain, 2020). Further, Baker & Smith (2019) approached educational AI tools from three different perspectives. The first one is about learner-facing, the second one is about teacher-facing, and the third one is about system-facing Artificial Intelligence in Education. Teacher-facing systems are used to support the teacher and reduce his or her workload by automating tasks such as administration, assessment, feedback, and plagiarism detection (Baker & Smith, 2019).

Ahmad et al. (2022) conducted research and found multifarious AI-based grading tools for higher education. Among them were “Write to learn”, “Quantum”, “Azure AI Edu”, “Hubert.AI”, “LightSide”, “Proctorio”, as well as “GradeScope” and “Respondus”. Most of these have a focus on facilitating the educator's work in terms of ID identification, evaluation, interpretation, and grading process. Additionally, Ahmad et al (2022) found that some of these tools are able to provide feedback to students in order to improve their learning experience as well as their development process.

Although the perceptions of students of being assessed by AI-based tools were already being researched by Sanchez-Prieto et al (2020), a very important aspect not covered by existing literature is the preference of students regarding which part of the grading process should be covered by AI and which by teachers as well as the students´ trust in AI-based assessment tools. Various scientific papers exist covering the general occurrence of AI in higher education or that AI supports teachers, but the opinions of students are not being considered. This represents the gap this paper aims to close in literature. By receiving and analyzing the answers of students, important insights can be published and used in science. The added value of this paper is the identification of challenges and opportunities for AI-based grading tools for open exam questions in higher education with the focus lying on the human part in the loop. The acquired knowledge could thus also be used to add more value and insights to the future development of the EasyGrader tool, especially for the question of where AI should be active in the grading process and where humans are preferred, from the point of view of students.

This will be realized by considering the precious opinion of students. Last but not least, one should always keep in mind that AI systems require control by humans. Even the smartest AI systems can make mistakes. AI systems are only as smart as the last date used to train them (Kaplan & Haenlein, 2019). This statement supports the idea of considering the opinion of students and what they think of AI´s role in grading open exam questions in higher education. By considering their opinions, the gap in literature can be closed. Also, this topic is important since it potentially affects the students´ professional future careers.

6. THEORETICAL FRAMEWORK

When taking a closer look at the research question, “What is the opinion of students regarding which steps the AI-based-grading tool “EasyGrader” in higher education should occupy in grading open exam questions and which step should be done by teachers?”, it is advantageous to clarify some of the terms and concepts.

6.1.1 Artificial Intelligence

Firstly, the term Artificial Intelligence (AI) needs to be defined. Numerous definitions exist, among the most popular are the following:

- Artificial intelligence is the mechanical simulation of collecting knowledge and information and processing intelligence of the universe: (collating and interpreting) and disseminating it to the election in the form of an actionable intelligence (Grewal, 2014)
- Artificial Intelligence is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of AI include expert systems, natural language processing, speech recognition, and machine vision (Burns, Laskowski, & Tucci, 2022)
- Computers that perform cognitive tasks are usually associated with human minds, particularly in the learning and problem-solving (Baker & Smith, 2019)

Since this research deals with the AI-based grading tool EasyGrader, which is able to execute tasks previously done by humans, the best fitting definition of AI for this paper is the one from Baker & Smith. The AI is executing the cognitive tasks needed, so the definition matches the given situation.

6.1.2 The grading process

Secondly, the framework of the different steps in grading needs to be specified. At the University level, grading mostly consists of the following steps: The first examiner checks every question and adds all points together. Afterward, the second examiner goes through the same process before both examiners compare both of their scores and determine the final score.

6.1.3 Open questions in exams

Thirdly, the concept of “open questions in exams in higher education” needs to be consolidated. “Open questions” in time-limited exams are questions that ask students to explain a certain process, model, or concept in their own words and written form. Open questions can be so-called fill-in-the-blank questions, short answers without a bank, or long answers.

7. RESEARCH METHOD/DESIGN

This research aims to systematically describe the acceptance among students regarding the use of AI-based tools for the grading of open questions in exams in higher education. The research method/design chosen for this study is a survey research design. This study was best fitting a survey since the goal was to get to a qualitative and interpretative instead of a quantitative and statistically based result. Only a minority of questions with Likert scale answers gave quantitative answers. The data needed to be collected primarily, meaning by the researcher. Respondents have been exposed to several questions and statements with different answering possibilities, among which are open questions, the Likert scale, and multiple-choice. Since the respondents gave their consent to take the survey and little to no private or personal data was asked (only the type of study was asked) and no data at all was stored, there was no ethical issue with this research method. The validity and reliability of this survey are insured through the limitation of only sending the survey to students since they are the main target of the research question.

8. SURVEY QUESTIONS/STATEMENTS AND HYPOTHESES/ASSUMPTIONS

The survey consists of 13 questions/statements/. Behind the questions/statements, there are certain hypotheses and assumptions, which were tested by asking the related questions and presenting the statements. Since the hypotheses did exist before analyzing the survey questions, this paper used an inductive approach to the thematic analyses. In the following, the questions and statements, as well as their belonging hypotheses and assumptions, are presented. In order to explain the hypothesis behind the questions, the hypothesis will be derived as well. *Q* will represent the questions and statements, and *H* will represent the hypotheses. Since some of the questions refer to the same hypothesis, not every question has a hypothesis but an assumption instead. Further, assumptions are being used for open questions where there is more room for interpretation and hypotheses are being used for multiple-choice questions where there are quantitative answers. Assumptions will be displayed as *A*.

Q1: What exactly do you study

H1: Students from Business/Computer science/IT studies are better informed about AI than students from other studies. This hypothesis is derived from the conjecture that students from studies related to the likes of Business, Computer Science, and Information Technology might have a better understanding of what Artificial Intelligence is since they are being or have been confronted with the topic throughout their study program compared to students from social sciences, health, design-related or different studies who might not be as informed about that specific topic. This hypothesis is connected to *Q1, Q2,* and *Q3.*

Q2: Please react to the following statement by using the Likert scale: I have the feeling that I understand what Artificial Intelligence is and that I grasp the concept behind it

A2: If a student has heard of AI, it is mostly very general knowledge and no deeper expertise. The topic of AI has become a very popular field in technology and many other areas of human life (Pannu, 2015). Nevertheless, to understand the topic and grasp the concept behind it, some kind of expertise, experience, deeper understanding, or application of AI in a work or study-program environment is needed. That is why probably a large share of respondents have heard about AI and know somehow how it works, but are not sure how exactly machine learning and algorithms operate.

Q3: Do you know that Artificial Intelligence-based tools exist/are being developed for grading open questions of exams in higher education?

A3: Most students do not know about the development of AI-based grading tools. It can be assumed that only little to no respondents know that AI-based grading tools are being developed/exist already since this technology is relatively young and generally unknown. A reason for this might be that the opinion or acceptance of students regarding AI-based grading tools for open questions in higher education has not been researched yet. Only those who know more about AI, in general, might know about the development/existence of such tools.

Q4: Which Artificial-Intelligence based tools for grading in higher education do you know?

H4: Most students do not know any AI-based grading tools. Building on the reasoning of *A3*, it is most probable that if most students do not know about the development and/or existence of AI-based grading tools, they also do not know or can not name any already existing AI-based grading tool.

Q5: Do you think Artificial Intelligence should only support teachers when grading exams in higher education but not do it alone (no teachers´ surveillance)? Please justify your answer!
A5: Most students would only want AI for supporting teachers. Although there is increasing trust in Artificial Intelligence in education in general Rutner & Scott (2022), only little trust can be expected in AI-based grading tools for open exam questions, since the shift from teacher-based grading to AI-based grading represents a sudden and big change in status quo of grading and thus a completely new situation for students. This is why students could be rather resistant to the change (Kotter & Schlesinger, 1989).

Q6: Do you think Artificial Intelligence-based tools should be part of the grading process of exams in higher education?
H6: Some students want AI to be part of the grading process, others do not (divided opinion). Again, referring the Kotter & Schlesinger (1989) and their theory of resistance to change, it might well be the case that some students do not want AI to be part of the grading process just because it represents something entirely new and/or unknown to them. Nevertheless, others might see the benefits AI can bring to the table and are in favor of implementing AI.

Q7: Would you rather want your next exam being graded by Artificial Intelligence only, human (teacher) only, or a hybrid approach?
H7: Most students vote in favor of Human only, hybrid approach, Artificial Intelligence only (descending order). Repeatedly, students most probably will choose the option they are used to and will not choose for a change, which in this multiple-choice question is "Human only", followed by "hybrid approach", which will most probably also get several votes, followed by "Artificial Intelligence only" as the least popular answer since this approach represents the most radical change.

Q8: The parts of the grading process are: checking (first examiner), checking (second examiner), and adding the points for the final score (also think of other steps you know). If you choose for/can imagine a hybrid approach, which steps of the grading process of open questions should be occupied by humans (teacher) and which by Artificial Intelligence? Please justify your answer!
A8: Most students will choose a hybrid approach, meaning that AI does the first/second checking and teachers do the other one. When it comes to the design of the grading process and it is assumed students can imagine a hybrid approach, it is not unlikely they will assign AI to do one part of the grading and humans the other one, whilst the specific arrangement does not play an important role.

Q9: Can you imagine Artificial Intelligence-based tools grading multiple-choice questions only, open questions only, both, or none of them? (tick the box)
H9: Most students vote in favor of Multiple-choice only, both, none of them, open questions only (descending order). Based on the previous hypotheses/assumptions (*5-8*), it is not unlikely that most students will tick "Multiple-choice only" at this question. Following the line of this argumentation, the consequence would then be that "both", "none of them" and "open-questions only" will follow.

Q10: Can you imagine Artificial Intelligence grading open questions for exams in higher education? If yes, with the support of humans or not? Please explain why!
A10: Most students do not want AI to be grading the open questions without teachers´ supervision. Once again referring to the existence of general trust in AI (Rutner & Scott, 2022), but a lack of trust in AI-based grading because of missing data regarding students´ acceptance of AI, it can be expected that students will not be trusting AI grading open questions on its own but to a minimum include the supervision of teachers.

A11: Are there cases where you can imagine Artificial Intelligence doing the grading on its own? Please provide examples!
H11: If students can imagine AI doing a part of grading, it is multiple-choice questions. Since for the largest share of students, the existence of AI in grading will probably be a new cognizance, most of the respondents will likely answer that multiple-choice questions are a good fit for letting AI grade on its own since the perception might be that grading these kinds of questions will not be as complex and that AI cannot do a bad performance there.

Q12: Which part of grading do you think can only be occupied by humans (teachers) and not by Artificial Intelligence?

A12: Most students would say open questions when being asked which part only teachers should occupy. Probably trusting humans (teachers) more than AI, it is likely most students will say that only teachers should grade open questions.

Q13: Which part of grading do you think can only be occupied by Artificial Intelligence and not by humans (teachers)?
A13: If a student can imagine AI exclusively doing a part of grading, he/she will choose multiple choice questions or "none". Referring to *H11* again, students will most probably choose AI exclusively when grading multiple-choice questions, since trust in this new system is not yet established enough and multiple-choice questions are considered "easy to grade", even for AI. Also, AI is probably not being trusted to grade open questions. The other probable outcome will be that students just say "none" since they have no trust in AI at all.

8.1 Data collection

The sampling method used for this research was to send out the survey to students via mail and social media. To exclude any unreliable participants the survey was only sent to students. The survey was conducted between May 9 and May 23, 2022. The survey consisted of 13 questions and statements that the participants were asked to answer and reflect on. Answer possibilities were the Likert scale, multiple-choice and open answers. It took the participants approximately ten minutes to conduct the survey. For analyzing the data, non-probability sampling was used since the type of this research is based on quality instead of quantity. Although this approach might be more at risk for sampling bias, the sample is still representative of the whole population. The aimed sample size was 50 and the response rate was 43 in the end. Every question or statement of the questionnaire inherently had a hypothesis/assumption that was about to be tested. As for some hypotheses it was not only sufficient to rely on respondents´ answers, external existing data had to be gathered for the analysis. The material sourced for this was the "Use of Artificial Intelligence to Grade Student Discussion Boards: an explorative study" by Rutner & Scott (2022). The material was used for this research since it addresses the acceptance of Artificial Intelligence in grading exams. Furthermore, Pannu (2015) was considered because of his findings on AI in general, as well as Kotter & Schlesinger's (1989) findings on resistance to change were used to explain why students might react critically toward the implementation of AI-based grading tools.

8.2 Data analysis

In order to analyze the data, this research uses thematic analysis to explore the respondents´ views as well as opinions expressed in the answers to the survey. This approach allows for flexibility when interpreting the data and grouping it into broad themes. This approach requires a lot of attention and care when executing. The thematic analysis follows a deductive approach since the hypotheses behind every question are being tested with the survey results.

8.3 Validation of the design

The qualitative survey approach was used for this research since the versatility of the answers of respondents offers a lot of multifaceted impressions of the student's opinion on AI-based grading tools in higher education. Especially with open questions, this study tries to capture as many opinions as possible to come to a broad set of answers in order to contribute new knowledge about the acceptance and perception of AI in grading among students.

9. RESULTS

In the following, the results of the survey are displayed in the same order as in the survey. The survey consists of 13 questions/statements and has been answered by 43 respondents. The questions and statements are displayed in italics, the results in regular font.

Q1: What exactly do you study?

This open question was a required one, which means that all of the 43 respondents answered. Assigning the answers to certain groups based on the content of the study, with 15 responses International Business Administration represented the most popular study, followed by Marketing Management (6) and Betriebswirtschaftslehre (4; – the German version of Business Administration). Business Information Technology (2) also belongs to this group. Engineering was also among the answers; assigned to this group are Industrial Design Engineering (2) and Industrial and Mechanical Engineering (1). The Computer Science group is represented by Information technology (2). Answers from the teaching direction were German and Spanish for the teaching profession (1), Nutrition/Home Economics and sports for the teaching profession (1), and Philosophy and Economics for the teaching profession (1). The social group of studies consisted of Social Anthropology and Geography (1) and Social Work (1). The last group consists of different studies which, based on their little occurrence in the study, cannot be grouped in their own group. These are Law (1), Ecosystem Management (1), Management Society and Technology (1), Medicine (1), Sports Management (1), and Psychology (1).

Summary of Answers for Question 1, (Table 1)

Study	Count
Business Administration	27
Engineering	3
Computer Science	2
Teaching Profession	3
Social studies	2
Mixed Group	6

Q2: Please react to the following statement by using the Likert scale: I have the feeling that I understand what Artificial Intelligence is and that I grasp the concept behind it

This statement was answered by 43 respondents (100% response rate).

Summary of Answers for Question 2, (Table 2)

Answer Category	Count	Count in %
1 (Not at all)	0	0%
2	5	11.6%
3	15	34.9%
4	17	39.5%
5 (very much)	6	14%

Q3: Do you know that Artificial Intelligence-based tools exist/are being developed for grading open questions of exams in higher education?

This question was answered by 43 respondents (100% response rate).

Summary of Answers for Question 3, (Table 3)

Answer	Count	Count in %
No	32	74.4%
Yes	8	18.6%
No answer	3	7%

Q4: Which Artificial-Intelligence based tools for grading in higher education do you know?

This open question was answered by 27 respondents. 22 out of the 27 respondents did say they “do not know any” AI-based grading tools for higher education, two respondents said “multiple-choice maybe”, one respondent said that he “heard something about US-based Universities that use AI-based grading tools” but that he forgot the name, and one respondent answered with “Speedgrader”.

Q5: Do you think Artificial Intelligence should only support teachers when grading exams in higher education but not do it alone (no teachers´ surveillance)? Please justify your answer!

The answers to this open question will be grouped into a set of categories, based on their content. 36 respondents answered this question. In total, eleven answers said that AI should only support teachers in the grading process since it might overlook certain aspects of the answers. Five respondents said that AI is subject to making errors, so teachers should support the AI system. Nevertheless, they can imagine that AI can grade exams on its own in the future, once it is further developed. Three respondents said that AI should only serve as a support to teachers when the exam questions are about opinions since AI is not able to identify certain aspects.

Furthermore, four respondents said that AI should only support teachers because of safety reasons. Further, three respondents said that they would prefer a hybrid grading approach since both teachers and AI can make mistakes. One respondent noted that AI can even grade exams on its own, provided the system is being checked every once in a while. Also, one respondent said that Multiple-choice questions can be checked exclusively by AI but that theory-related questions should be supervised by teachers. Besides, one respondent said that AI should only support teachers based on efficiency reasons. Another respondent noted that AI might work in certain studies but not in others. He/she cannot imagine AI doing the grading in for example Anthropology, since the AI would not be able to identify cultural viewpoints. Further, one respondent said that grading would be more reliable if only real people do it. Similar to this, two respondents said that they do not trust AI in grading exams at all. Contrary to that, one respondent noted that AI is fairer in grading compared to teachers since teachers have prejudices and sympathies/antipathies which would be eliminated through AI. Three responses were off-topic.

Q6: Do you think Artificial Intelligence-based tools should be part of the grading process of exams in higher education?

This Multiple-choice question got 43 responses (100% response rate).

Summary of Question 6, (Table 4)

Answer	Count	Count in %
Yes	37	86%
No	6	14%

Q7: Would you rather want your next exam being graded by Artificial Intelligence only, human (teacher) only, or a hybrid approach?

This Multiple-choice question was answered by 43 respondents (100% response rate).

Summary of Question 7, (Table 5)

Answer	Count	Count in %
Hybrid approach	37	86%
Human (teacher) only	5	13.5%
Artificial Intelligence only	1	0.5%

Q8: The parts of the grading process are: checking (first examiner), checking (second examiner), and adding the points for the final score (also think of other steps you know). If you choose for/can imagine a hybrid approach, which steps of the grading process of open questions should be occupied by humans (teacher) and which by Artificial Intelligence? Please justify your answer!

This open question was answered by 31 respondents. The answers will be assigned to categories based on their content. 17 respondents said that AI should do the first checking, teachers should do the second checking and AI should calculate the score. One of these 17 respondents named time efficiency as a reason, another one named possible AI mistakes, and a third one said that AI should start grading so that teachers are unbiased. A different respondent said that AI should do the first checking, and teachers the second checking as well as add the scores for them to correct possible AI mistakes. Another respondent said that AI should do both checking steps and that teachers afterward check for AI mistakes and calculate the score.

Furthermore, one respondent said that AI should do the first checking, teachers the second checking, and both AI and teachers to calculate the score. Also, one respondent said that AI should do the first checking, teachers the second checking, and students to review both checking processes. Further, one respondent said that teachers should do the first checking and AI the second checking to make sure the teachers did not make any mistakes. Additionally, one respondent answered that it doesn't matter whether AI or teachers do the first and/or second checking and that a hybrid approach makes the most sense. Another respondent said that both parties should do both checking processes. Two respondents said that teachers should do the first checking, AI the second checking, and teachers calculate the final score. Furthermore, one respondent said that AI should only act as a third inspector to resolve human (teacher)-made mistakes. Further, one respondent said that AI should only scan for keywords. One respondent said that AI should only calculate the final score, while another one said that AI should not grade open questions at all. One answer was off-topic.

Q9: Can you imagine Artificial Intelligence-based tools grading multiple-choice questions only, open questions only, both, or none of them? (tick the box)

This Multiple-choice answer got 43 responses (100% response rate).

Summary of Question 9, (Table 6)

Answer	Count	Count in %
Both	26	60.5%
Multiple-choice only	17	39.5%
Open questions only	0	0%
None of them	0	0%

Q10: Can you imagine Artificial Intelligence grading open questions for exams in higher education? If yes, with the support of humans or not? Please explain why!

This open question was answered by 36 respondents. The answers will be grouped into categories based on their content. The most common answer (15 respondents) was that AI can grade open exam questions but only with the support/supervision of teachers. A group of three respondents said that AI can grade open questions since it has less room for interpretation and thus offers a more uniform assessment. Nevertheless, a teacher should check the AI-based grading in the end. Another group of three respondents thinks AI can grade open questions with a good quality keyword scanning which, nevertheless, should be checked and supervised by teachers in the end. Further, one respondent said that AI can grade open questions, but that the data should be checked oftentimes so that the quality of the AI tool is ensured and maintained, while another respondent can imagine AI grading objective questions but leaving subjective questions for the teachers. Related to this, one respondent said that AI should grade easy questions but leave complex questions for the teacher. Also, a different respondent said that AI should only grade with the support of teachers at first and once the tool is commonly seen as reliable, it should grade on its own. Another respondent said that a hybrid approach would be the best approach in order to cancel out the flaws of both parties.

Contrary to this, six respondents cannot imagine AI grading open exam questions at all. One out of these six said he thinks so because AI is not there yet, and another one said AI cannot identify and assess reasoning structures. A group of three people can only imagine humans grading open questions with the support of AI since the system is not yet developed well enough to operate on its own and that it potentially endangers the student's future. One respondent said that he does not know an answer to this question.

Q11: Are there cases where you can imagine Artificial Intelligence doing the grading on its own? Please provide examples!

This open question got answered by 37 respondents. The answers will be grouped into categories based on their content. The most popular answer (19 respondents) was that AI can grade Multiple-choice questions. A few of these respondents further claimed that AI can also grade single-choice questions, while another one said AI should only grade Multiple-choice questions if the examined student wishes so, while again another one added that AI should be grading Theses for correct citation and plagiarism. A further member of this group also can imagine AI to grade grammar-based questions. A group of six respondents said that AI should grade Multiple-choice questions as well as questions where the answer is known, meaning a black and white answer possibility. Much the same as this, three respondents said that AI should grade open questions as well as number-based ones. Two other respondents said that AI can be used to grade questions where theory is asked; one of them said AI could be used for grading listening comprehension. Another respondent said that if the result of the exam is not essential, AI should do the grading, while a different respondent mentioned that AI should only do the grading in the future. Furthermore, one respondent stated the opinion that once AI is commonly seen as reliable, it should grade open questions of exams, so that students would see their results immediately after submitting their exam. In total, three respondents do not see AI doing any grading at all. One answer was off-topic.

Q12: Which part of grading do you think can only be occupied by humans (teachers) and not by Artificial Intelligence?

This open question was answered by 34 respondents. The answers will be grouped into categories based on their content. Here, the most frequent answer (seven respondents) was that only teachers should grade open questions, where there is much room for interpretation. One example named was philosophy exams. Just as popular (seven respondents) was the answer that teachers should grade open exam questions in general. Further, six respondents said that teachers should grade questions with subjective answers, two of which mentioned that also abstract answers should be graded that way. Three respondents shared the opinion that teachers should calculate the final score, one of them also wanted teachers to review the whole exam. A pair of two students were convinced teachers should grade reports. Two respondents mentioned that teachers should grade feedback-related questions as well as answers with long text. Further, one respondent said only teachers should grade exams when the result is really important. Additionally, one respondent mentioned that for ethical and logical questions teachers should be the grading party. In opposite to the aforementioned answers, two respondents said that AI should do the whole grading process with no teacher-based support. Comparable to this, one respondent said AI should do the whole grading process but teachers should be present at the exam review. Two answers were off-topic.

Q13: Which part of grading do you think can only be occupied by Artificial Intelligence and not by humans (teachers)?

This open question was answered by 30 respondents. The answers will be grouped into categories based on their content. The most popular answer (eight respondents) was that only AI should grade Multiple-choice questions. One of these eight respondents also named math problems, another one said single-choice questions as well and a third one mentioned that AI can also grade open questions.

Two respondents said that AI should grade big paragraphs since it is able to analyze those much more quickly than teachers. One of them also mentioned that AI should calculate the final score, while a different respondent only wants AI to do the scores. Further, another respondent said that AI should grade speed and accuracy tests. Two respondents were convinced that teachers, as well as the AI, can both do every part of grading, but that AI would be more efficient. Another pair of two respondents said that everything can and should be done by teachers, while a different respondent said everything should be done by the AI. Two further respondents said AI should only do fraud checking, while another one said AI should only deliver the first overview so that the teacher can grade in the end.

A large share of seven respondents did not want AI to grade anything on its own at all. Three respondents failed to answer the question by missing the topic.

10. DISCUSSION

In this section, the results will be interpreted by also highlighting the hypotheses and whether these turn out to be true or false. Since this study uses a qualitative approach, no statistical significance will be used. Instead, interpretations of the results will be conducted. The same order of Hypotheses/questions from section 8, Survey questions/statements and Hypotheses/Assumptions, will be used.

H1: Students from Business/Computer science/IT studies are better informed about AI than students from other studies. After reviewing the results of Q1 in connection with Q2 and Q3, where this research has measured the average answer of Business/computer science/IT students to be 3.7 on the Likert scale regarding knowledge of AI and the average answer of all other studies being 3.3 (results Q2), this hypothesis will be accepted. This implies that students from Business, computer science, or IT-related studies are better informed about AI in general than students from all other studies.

A2: If a student has heard of AI, it is mostly very general knowledge and no deeper expertise. The result of this question varies. Besides "1" on the Likert scale, meaning that no knowledge about AI is given, all other answer possibilities have been chosen (2-5). With a mean value of 3.5, it appears that most students have a general knowledge of AI but it seems to be questionable that a lot of respondents have deeper expertise since only six out of 43 respondents claimed that they completely grasp the concept of AI ("5" on the Likert scale), which implies that this assumption can be accepted and that students (the respondents) in general do not have deeper expertise in AI.

A3: Most students do not know about the development of AI-based grading tools. When asked after knowing about AI-based grading tools being developed, 32 out of 43 respondents answered with "no". Three respondents chose "no answer" and at least 8 knew about the development of such tools. That means that the biggest share of respondents did not know about the development of AI-based grading tools, which appears to be the confirmation of this assumption.

H4: Most students do not know any AI-based grading tools. This hypothesis can be accepted since 26 out of the 27 respondents could not name any AI-based grading tool and only one respondent answered with "Speedgrader". This result implies that most students indeed do not know any AI-based grading tool, which could mean that this topic needs some more publicity to reach the target group (students) in order to enhance acceptance of the AI tools.

A5: Most students would only want AI for supporting teachers. Although a few respondents mentioned that in the future, AI can grade exams on its own, the majority of respondents said that AI still needs the support of teachers, which implies that the trust and acceptance of AI are yet to be consolidated among students. Furthermore, a few respondents did not want AI to grade exams at all, which supports this implication. That is why the assumption can be accepted.

H6: Some students want AI to be part of the grading process, others do not (divided opinion). This hypothesis can be seen as false. Since 37 out of 43 respondents want AI to be part of the grading process in higher education and only six respondents do not, the biggest share of students appears to be not resistant to change, different than assumed before, and open-minded regarding the implementation of AI in grading.

H7: Most students vote in favor of Human only, hybrid approach, Artificial Intelligence only (descending order). This hypothesis again can be seen as false. Out of 43 respondents, 37 respondents ticked "hybrid approach", five respondents choose "human (teacher) only" and one respondent ticked "Artificial Intelligence only". This result implies that the most favored approach for the future of grading in higher education among students is a hybrid approach of AI and teachers.

A8: Most students will choose a hybrid approach, meaning that AI does the first/second checking and humans do the other one. Out of 31 respondents, 17 respondents said that AI should do the first checking, teachers should do the second checking and AI should calculate the score. The overall consensus was that AI occupies one part of the checking, while teachers do another one, which implies that a hybrid approach is most popular among students, which in turn means that this assumption can be accepted.

H9: Most students vote in favor of Multiple-choice only, both, none of them, open questions only (descending order). Out of 43 responses, 26 respondents ticked "both", 17 respondents ticked "Multiple-choice only", while "open questions only" and "none of them" both got no votes at all. This result appears to be evidence that the hypothesis is false and that students can also imagine AI to grade open questions as well.

A10: Most students do not want AI to be grading the open question without teachers´ supervision. The results show that approximately half of all respondents said that AI can grade open questions but only with the supervision of teachers, which implies that this assumption can be accepted. This implies that for the further development of AI-based grading tools, teachers should always be considered.

A11: If students can imagine AI doing a part of grading, it is multiple-choice questions. The biggest share of respondents was convinced that AI should only grade multiple-choice or single-choice questions, which appears to be evidence that this assumption can be accepted. Here it appears to be the case that most students do not want open questions to be graded by the AI.

A12: Most students would say open questions when being asked which part only teachers should occupy. Also based on the result of the previous assumption, most students here for this question actually want teachers to grade open questions and not the AI, which implies that this assumption can be accepted.

A13: If a student can imagine AI exclusively doing a part of grading, he/she will choose multiple choice questions or "none". Since the results of this aspect show a lot of variation, this assumption cannot be accepted and is thus wrong. One of the most popular answers here was that AI should grade multiple-choice questions. Curiously enough, among the most popular answers is also that AI should not grade anything at all and that only AI should grade open questions, which implies that there is not a common opinion on AI in grading and that acceptance of AI is not omnipresent.

10.1 Limitations

This paragraph will display the limitations of this research. It has to be mentioned, that the sample size, as well as the sampling method, are subject to limitations. Since this research mostly offers the opinions and answers of people known to the researcher, the sample might not be as independent. Especially, since a lot of respondents stem from business-related studies and live in either Germany, the Netherlands, or Austria, which might be resulting in a warped or unilateral displaying of results. Furthermore, the study is limited to time and scope, since the time granted for the realization of the study from the University side was rather short.

11. CONCLUSION

This research aimed to identify students´ opinions on AI-based grading tools for open exam questions in higher education. Based on a quantitative analysis of students´ answers to a set of questions/statements regarding the aforementioned topic, it can be concluded that students, in general, are not averse to the implementation of AI for grading purposes in higher education. Important to mention is, that most of them can imagine AI only grading multiple-choice questions. A large share can also imagine AI to grade open exam questions, but only with the supervision of teachers. Furthermore, only a few students knew that AI-based grading tools exist or are being developed and only one student could name such a tool. Another crucial aspect is that the knowledge of AI and how it works, in general, is not excessively well known by every student and the acceptance of AI in grading in higher education is not commonly agreed on among students, which implies that there needs to be more information and education in order to create awareness, acceptance, and trust of the potential new grading through AI. To better understand the implications of these results, future studies could aim to find out why the acceptance of AI in higher education among students differs depending on their studies and/or other circumstances. Returning to the problem statement, namely, that the perceptions of students regarding their assessment through AI-based tools were already being researched by Sanchez-Prieto et al (2020) but a very important aspect not covered by existing literature is the preference of students regarding which part of the grading process should be covered by AI and which by teachers as well as the students´ trust in AI-based assessment tools, this research addressed the aforementioned gap in the literature. This research has shown that students mainly want AI to only grade multiple-choice questions on its own and if they can imagine AI also grading open questions, a teacher has to stay in place and supervise it. Another popular opinion is that only teachers should grade open questions and also a minority does not want AI to participate in grading at all. These results show a general acceptance of AI in grading in higher education but a divided opinion on whether open questions should be graded by AI, which implies that the acceptance of AI and thus also for the AI-based grading tool "Speedgrader" in grading in higher education among students is subject to expansion in the future.

11.1 Recommendations for future research

A proposition for future research would be to execute similar research in another environment, meaning another context, country, or culture, and with a geographically as well as study-related more dispersed sample of respondents. Furthermore, a larger time frame and scope might lead to new or more exclusive insights or new knowledge. Also, the re-assessment and broadening of theory might be a suggestion for future research.

12. ACKNOWLEDGMENTS

For the completion of this paper cooperation with several parties and the support of supervisors as well as friends and family were needed. I would like to thank Dr. Daniel Braun and Dr. Patricia Rogetzer for their support and helpfulness. Further, I would like to say thank you to all survey respondents. Most importantly, I would like to thank my Father, Mother, Brother, Girlfriend, and closest friends circle for supporting and encouraging me and stimulating my creativity.

13. REFERENCES

Ahmad, K., Maabreh, M., Ghaly, M., Khan, K., Qadir, J., & Al-Fuqaha, A. (2022). Developing future human-centered smart cities: Critical analysis of smart city security, Data management, and Ethical challenges. *Computer Science Review*, *43*, 100452.

Baker, T., & Smith, L. (2019). Educ-AI-tion rebooted? Exploring the future of artificial intelligence in schools and colleges. Retrieved from Nesta Foundation website: https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.pdf

Brennen, E., 2020, AI powered grading software earns high marks, https://www.uml.edu/news/stories/2020/gradescope-software.aspx

Burns E., Laskowski N., Tucci L. (2022) What is Artificial Intelligence (AI)? *TechTarget.com*, last updated February 2022

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *Ieee Access*, *8*, 75264-75278.

Clarke, V., Braun, V., & Hayfield, N. (2015). Thematic analysis. *Qualitative psychology: A practical guide to research methods*, *222*, 248.

Grewal, D. S. (2014). A critical conceptual analysis of definitions of artificial intelligence as applicable to computer engineering. *IOSR Journal of Computer Engineering*, *16*(2), 9-13.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004.

Kersting, N. B., Sherin, B. L., & Stigler, J. W. (2014). Automated scoring of teachers' open-ended responses to video prompts: Bringing the classroom-video-analysis assessment to scale. *Educational and Psychological Measurement,74*(6), 950–974. https://doi-org.ezproxy2.utwente.nl/10.1177/0013164414521634

Kotter, J. P., & Schlesinger, L. A. (1989). Choosing strategies for change. *Readings in strategic management*, *1*, 294-306.

Meyer, J. W., Ramirez, F. O., Frank, D. J., & Schofer, E. (2007). Higher education as an institution. *Sociology of higher education: Contributions and their contexts*, *187*.

Moeini, M., Rahrovani, Y., & Chan, Y. E. (2019). A review of the practical relevance of IS strategy scholarly research. *The Journal of Strategic Information Systems*, *28*(2), 196-217.

Pannu, A. (2015). Artificial intelligence and its application in different areas. *Artificial Intelligence*, *4*(10), 79-84.

Rutner, S., & Scott, R. (2022). Use of Artificial Intelligence to Grade Student Discussion Boards: An Exploratory Study. *ISEDJ*, *20*(4), 4.

Rutner, S., & Scott, R. (2022). Use of Artificial Intelligence to Grade Student Discussion Boards: An Exploratory Study. *INFORMATION SYSTEMS EDUCATION JOURNAL*, *20*(4), 4.

Sánchez-Prieto, J. C., Cruz-Benito, J., Therón Sánchez, R., & García Peñalvo, F. J. (2020). Assessed by machines: development of a TAM-based tool to measure AI-based assessment acceptance among students. *International Journal of Interactive Multimedia and Artificial Intelligence*, *6*(4), 80.

Toffel, M. W. (2016). Enhancing the practical relevance of research. *Production and Operations Management*, *25*(9), 1493-1505.

Vij, S., Tayal, D., & Jain, A. (2020). A machine learning approach for automated evaluation of short answers using text similarity based on WordNet graphs. *Wireless Personal Communications*, *111*(2), 1271–1282. https://doi-org.ezproxy2.utwente.nl/10.1007/s11277-019-06913-x

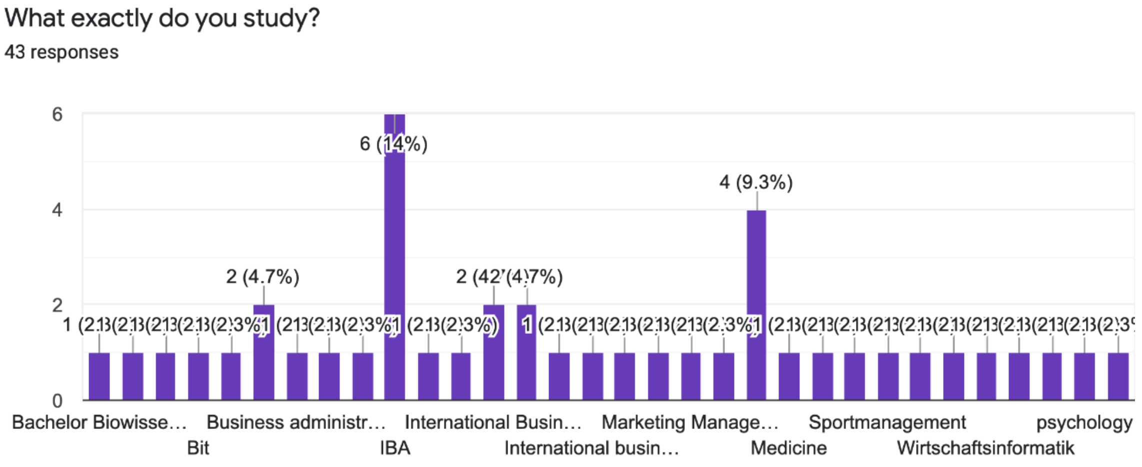
14. APPENDIX

Survey (created and spread via google forms)

The results that show graphs and output are copied from google docs and will be displayed here (Q1, Q2, Q3, Q4, Q6, Q7, Q9).

Open questions, which all have 30+ answers (Q5, Q8, Q10, Q11, Q12, Q13), can be reviewed following this link: https://docs.google.com/forms/d/1mAuaF8AsJBlL0zQxJhoYHXjvzLq2Xybyh8jQeN-PdfA/edit - responses

Q1)



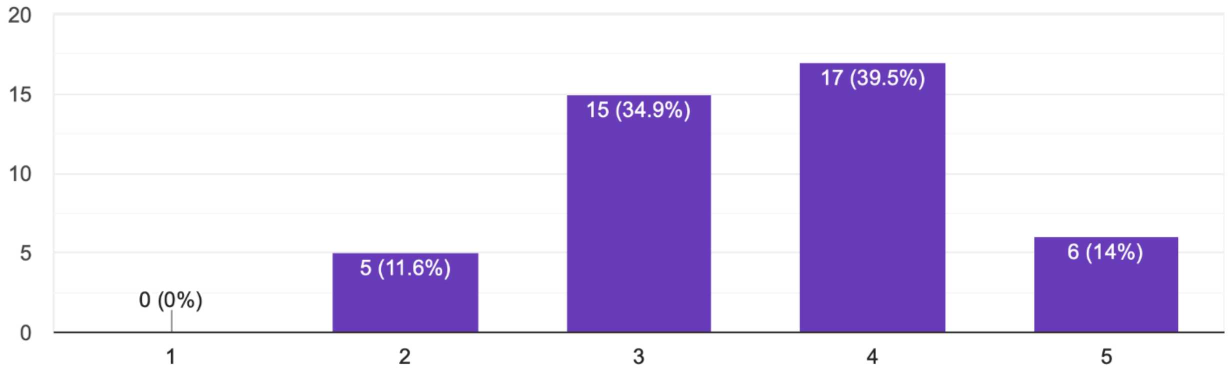
What exactly do you study?

43 responses

Q2)

Please react to the following statement using the Likert scale: I have the feeling that I understand what Artificial Intelligence is and that I grasp the concept behind it

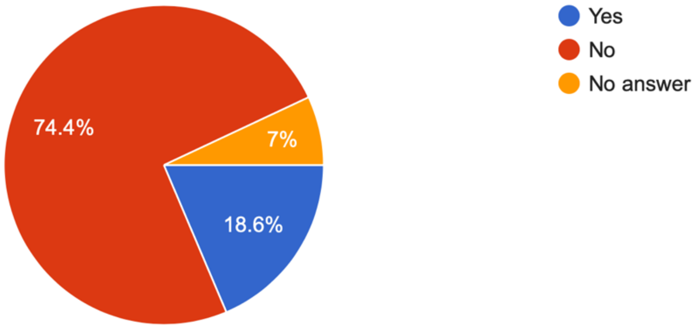
43 responses



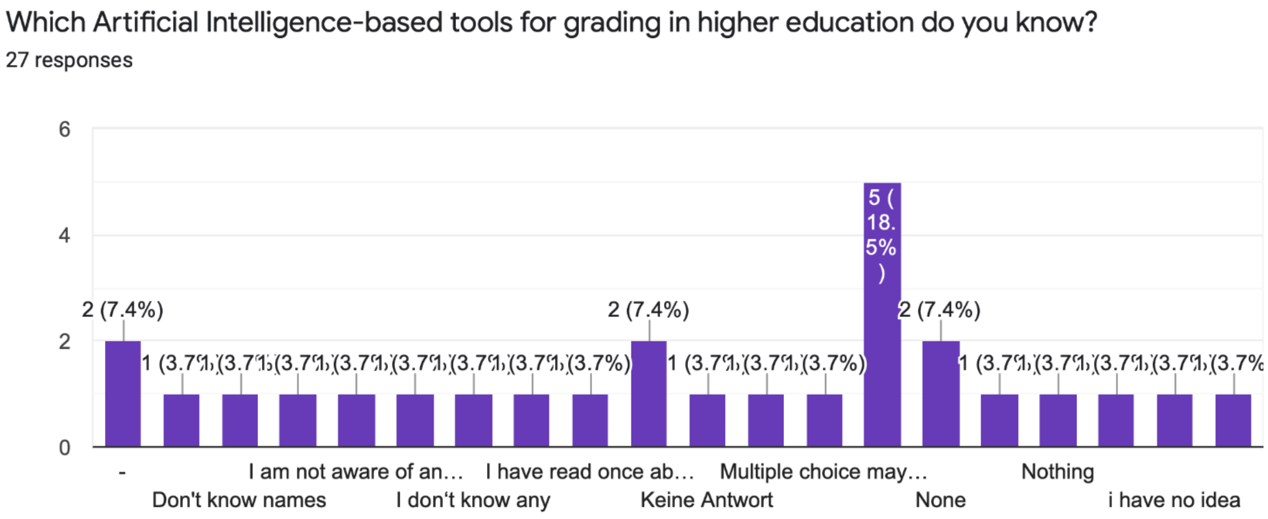
Q3)

Do you know that Artificial Intelligence-based tools exist/are being developed for grading open questions of exams in higher education?

43 responses



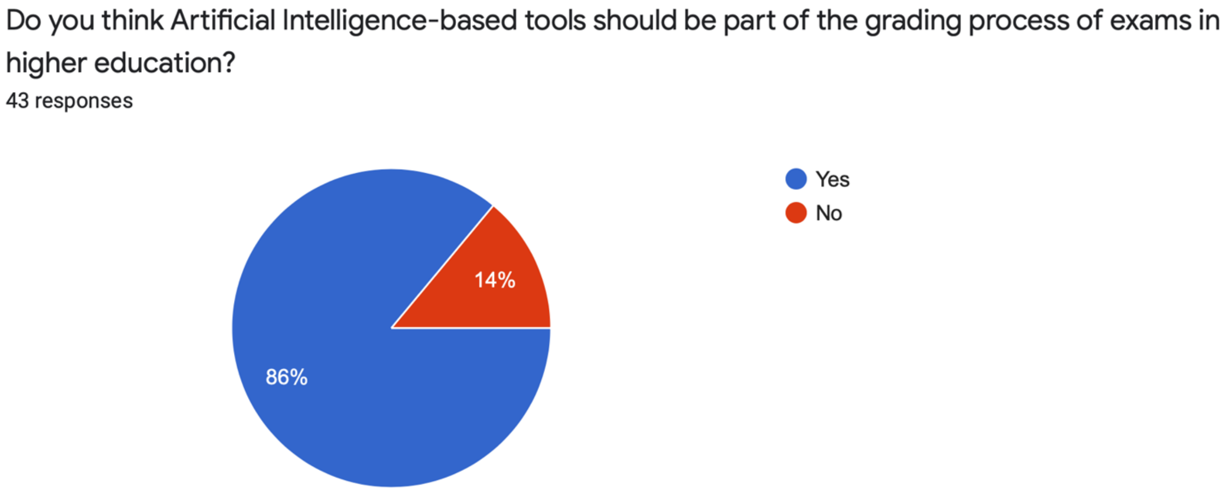
Q4)



Which Artificial Intelligence-based tools for grading in higher education do you know?

27 responses

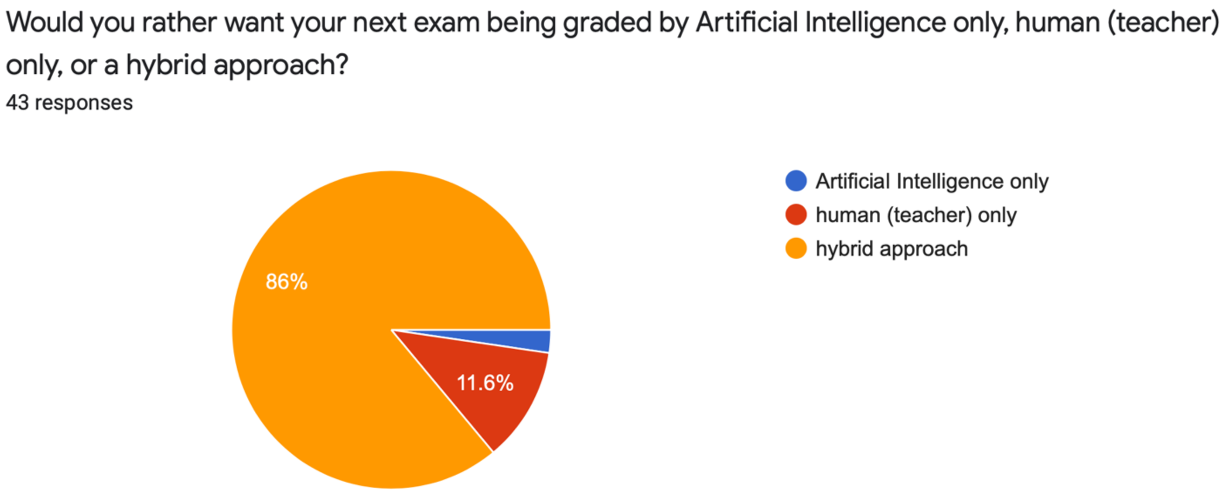
Q6)



Do you think Artificial Intelligence-based tools should be part of the grading process of exams in higher education?

43 responses

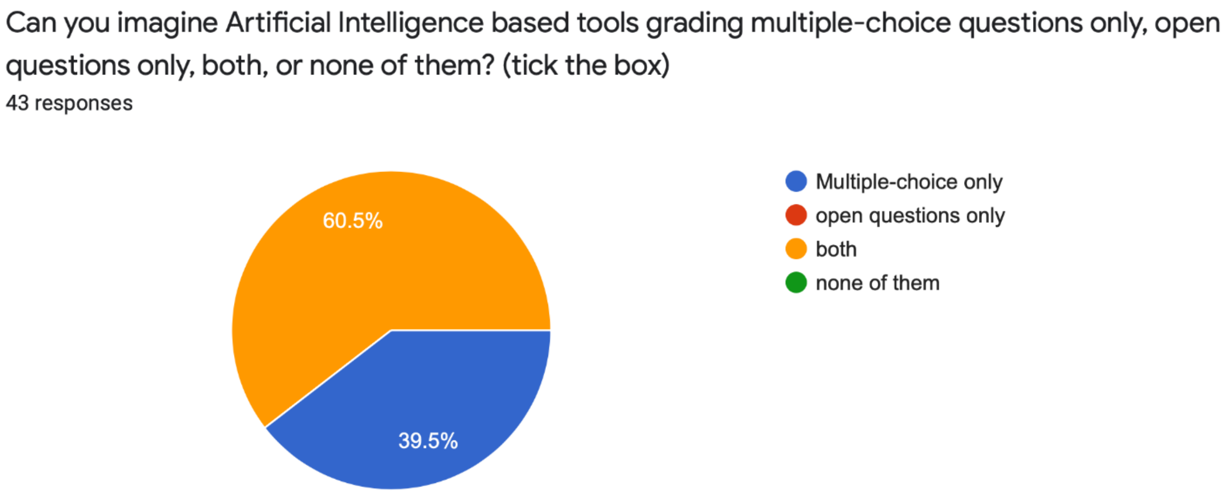
Q7)



Would you rather want your next exam being graded by Artificial Intelligence only, human (teacher) only, or a hybrid approach?

43 responses

Q9)



Can you imagine Artificial Intelligence based tools grading multiple-choice questions only, open questions only, both, or none of them? (tick the box)

43 responses