A Student's Take on Challenges of AI-driven Grading in Higher Education

Eva Stoica
University of Twente
PO Box 217, 7500 AE Enschede
the Netherlands
e.stoica@student.utwente.nl

ABSTRACT

Artificial Intelligence (AI) is one of the emerging innovations that are continuously shaping today's world. The usage and applications of AI can be observed in numerous fields, from medicine to education. However, when looking at these areas, there are still challenges that need to be overcome for the seamless integration of AI. For higher education, assessing students' capabilities is a critical factor in the learning and growing process. The grading procedure, especially for open-ended questions, can prove to be difficult and time-consuming for examiners. Hence, it is believed that AI might help in the grading of these types of questions. This research aims to discover the challenges that can arise with using AI for automated grading, from a teaching assistant's (TA) perspective. In the context of the research, a TA represents a student supporting educational activities. Furthermore, an overview of possible prerequisites that should be considered throughout the automated grading process is offered. This is done through literature reviews and by interviews and a survey with TAs in higher education. Whilst AI can be perceived differently by TAs, the results highlight that it represents the future. However, AI should only be used as a support tool for assessors and continuously improved for better acceptance and integration. The research brings value to an ongoing project focused on the development of an AI-supported grading tool at the University of Twente. Additionally, it adds information about automated grading to the scientific body of knowledge.

KEYWORDS

Artificial Intelligence, Automated Grading, Open-Ended Questions, Higher Education, Teaching Assistants

1 INTRODUCTION

As AI is advancing at a fast pace it is important to take advantage of the benefits that it can bring to a certain domain.

In education, the assessment of a student's performance plays a critical role in the forming process [1]. Nowadays, the number of students is increasing and performing assessments by teachers is becoming a time-consuming procedure that can lead to unreliable and biased results.

Currently, evaluating the gained knowledge of students (summative assessment) and stimulating the learning process (formative assessment) using computer-based systems is predominately done for multiple-choice questions. Nevertheless, in higher education, open questions, or short-answer questions have proven to produce better learning outcomes [26]. For the research, an open(-ended) or short(-answer) question asks for the student's input or opinions, in form of textual answers (e.g., reflective answers where a student uses his own words). These types of questions enhance meaningful learning by making students connect previously accumulated information, synthesize it, and present it as an application to a problem, or in short "recall, organize and integrate ideas" [13].

Considering the importance of open questions, potential applications of AI can help in automated grading. By having automated techniques to aid with assessments, faster feedback can be delivered. This can make the students improve quicker. Some challenges of AI in grading open questions can be the "relevance of the content to the prompt, development of ideas, cohesion and coherence" [22].

This research includes a literature review related to finding the possible challenges that arise for AI-driven grading. Additionally, interviews and a survey have been conducted with TAs from the University of Twente, to understand their attitudes towards automated grading. Teaching assistants represent the bridge between a teacher and a student. On one hand, they can be the ones responsible for (pre-)grading the open questions (teacher standpoint) and on the other hand, they can be the ones being graded (student perspective).

2 PROBLEM STATEMENT & RESEARCH QUESTION(S)

Assessing the performance of a student, by means of open questions, has a critical influence on the improvement and evaluation of the learning process. As the number of students increases, examiners find it more challenging to grade such questions, given that this is more time-consuming compared to the assessment of multiple-choice questions, and provide meaningful individual feedback. Here is where an application of AI would support the process.

From a student's perspective, the project aims to discover the conditions needed for AI to aid automated grading in higher education. This viewpoint is essential, considering that the assessment mostly impacts the students. For example, based on their answers, AI can be used, and the formative and summative assessments will influence their learning process.

The problem statement has led to the main research question that is investigated in this paper:

RQ – Under which prerequisites can Artificial Intelligence help the automated grading of open-ended questions in higher education?

Two sub-questions have been framed for a better work division from the main research question.

TScIT 37, July 8, 2022, Enschede, The Netherlands

RQ1 – What issues can arise whilst using Artificial Intelligence for the assessment of open-ended questions?

RQ2 – What is the TAs' attitude toward Artificial Intelligence for grading in higher education?

3 RELATED WORKS

A systematic search on scientific repositories was conducted to obtain a preliminary understanding of the prerequisites of AI in grading. The following domains were used: Scopus, ISI Web of Science, IEEE Xplore and ACM Digital Library.

The search started with the creation of an overview with essential keywords. Afterwards, search queries were used to retrieve relevant information, as seen in Table 1.

Table 1. Overview of Keywords, Search Queries and Search Domains

Keywords		*AI*	*Grading*	*Challenges*	*Open-Ended Question*	*Higher Education*	*Students*
Main Search Queries	*Scopus*	(Challeng* OR Prerequisit* OR Condition* OR Requirement) AND (AI OR Artificial Intelligence OR ML OR Machine Learning OR Automa*) AND (Grading OR Assessment* OR Examinat*)					
		((Open AND Question*) OR (Short AND Answer*)) AND (Grading OR Assessment OR Examinat*) AND (AI OR Artificial Intelligence OR ML OR Machine Learning OR Automa*)					
		(High* AND Education) AND (Grading OR Assessment OR Examinat*) AND (AI OR Artificial Intelligence OR ML OR Machine Learning OR Automa*)					
		(Student* OR Undergraduate) AND (Grading OR Assessment OR Examinat*) AND (AI OR Artificial Intelligence OR ML OR Machine Learning OR Automa*)					
	ISI Web of Science	automated open question grading					
		students' perspective on automated grading					
		grading machine learning short questions					
	ACM & IEEE Explore	automated grading higher education open questions					
		students' perspective on automated grading					
Note		*All the results were refined based on relevance, number of citations or the year.*					

From the initial searches, thirteen papers were reviewed, given that they were within the scope of the research. The useful references found in the initial papers have also been analysed. This led six other papers to bring new insight to the research.

The results of the existing literature, on approaching the problem, are described in the next subsections.

3.1 Challenges of AI in Grading

Numerous papers discuss the issues that can arise whilst using automated grading for open-ended questions. The automatic assessment of these types of questions remains an underused process [23]. Open or short answer questions have a great potential in higher education because they deepen the students' knowledge [14, 16]. Hence, it is important to understand why automated free text answer assessment remains unexploited.

One pressing concern is the limited feedback provided by automated grading systems [2, 7, 9, 11, 18, 27]. Feedback is crucial for meta-cognition, as it allows students to understand their mistakes and improve them. This way they can better communicate their knowledge, skills and understanding [18]. Furthermore, the consistency and the fairness of the feedback should be considered for automated grading. In higher education, teachers can have teaching assistants helping them in the assessment process, and this can generate differences in grading and feedback. Thus, having an AI-supported grading process to help with the examination can enable consistent and fair feedback [2]. Nonetheless, one can wonder how fair can an AI-supported tool be if this would be trained with the help of past exams. For example, if the training dataset will consider previously graded assignments, there might be cases in which an assessor made a mistake and graded a student incorrectly.

Another issue that stood out from the initial literature review was that no domain-specific knowledge is used for the automated grading [9, 22, 23]. This can become a challenge especially when different words have different meanings for particular subjects. If the context is not analysed, and an algorithm would look to see if a particular keyword exists in the answer, unfair results might appear.

Other mentioned challenges are the use of creative answers (case in which the solution would not match with the blueprint) [18, 29], not having the needed amount of data to (pre-)train models [9, 22] or the lack of reliability and validity of an AI tool [5, 18].

3.2 Techniques and Systems

The literature creates a picture of the possible techniques and systems that are used for automated grading, as well as their limitations and challenges. With continuous improvement and refinement, these tools can become an important aid for automated grading.

By far, the most mentioned techniques used for automated grading in education are Natural Language Processing (NLP) [2, 6, 9, 20, 22, 23, 27], followed by Machine Learning (ML) [12, 14, 22, 23]. Different approaches are describing how to properly integrate these techniques into beneficial systems. For example, NLP can detect errors in open-ended questions by comparing correct answers to the students' answers. Another discussed method is the "Bag of words". This represents a "graph-based method to find semantic similarities in short answer scoring" [22]. For this method, a sample is used to check the frequency of each word.

As for the existing systems, in two of the reviewed papers [3, 20], overviews are offered. The tools use various techniques, and each of them has benefits and drawbacks. This illustrates and supports the fact that the ideal implementation of AI in grading is yet to be created.

3.3 Students' Perspectives and Attitudes

Cases, studies, articles, surveys, and interviews including the students' take on automated grading revealed that there is limited information available in the literature focused on the students' attitude toward AI in grading.

However, the discovered papers have similar themes in common, when discussing the students' thoughts. For example, transparency and understanding of the way that the automated grading system is scoring the students were a reoccurring topic [5, 11, 18, 24]. It can be observed that technology acceptance is a highly influential factor in the way that AI is perceived by learners. Not trusting the automated grading process can even lead to increased levels of anxiety [14].

Most papers concluded that there are students who positively perceive AI grading tools, whilst others have negative thoughts, such as mistrust, that are disclosed.

Thus, this research addresses the existing gap in the literature related to the students' attitudes and perceptions regarding AI in grading. Moreover, new insights into possible challenges are investigated.

4 METHODOLOGY AND APPROACH

4.1 Selected Approach

This section describes the steps that were taken to perform the research. A qualitative research approach was selected. This type of research is believed to highlight the different perspectives of the participants and create a picture that portrays the existing challenges of automated grading in higher education. Therefore, qualitative research helps enhance the

"subjective meaning, action and context of those being researched" [10].

Figure 1 illustrates an overview of the research approach.

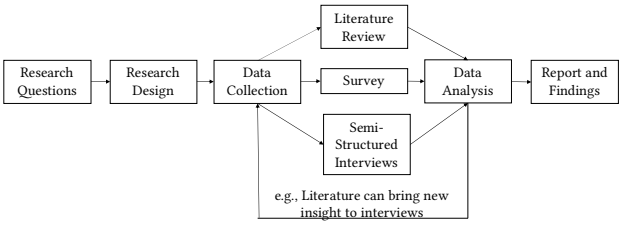

Figure 1. Research Approach

4.2 Importance of the Literature Review

A literature review was performed to find out what issues can arise with automated grading. Additionally, whilst doing the literature review, a search for documents expressing students' attitudes regarding AI in grading was conducted. The review is beneficial because it offers an understanding of the current results found in the literature. With the help of this process, potential answers to RQ1 are explored.

Furthermore, from the findings in the literature, the survey and the interview questions were refined. This was possible, on one hand, because the literature clearly highlighted "important" points that needed to be discussed with the future participants of the research. On the other hand, the briefly mentioned topics have also influenced the questions. In this particular case, it was interesting to observe the gap between the literature and a real context.

4.3 Design of the Interviews and Survey

Both a survey and semi-structured interviews were conducted for the collection of actual and relevant data on the topic. Surveys allow participants to answer the questions in their own time, anonymous. Thus, they yielded "unselfconscious data [...] in a short amount of time" [25]. Compared to surveys, interviews are valuable because they allow unexpected topics to surface, therefore offering a new angle on the research [4]. The participants were given the ability to choose their preferred option (either an interview or a survey), such that a broader understanding of the problems could be created.

The target population of the research is represented by students that are/have previously worked as teaching assistants in higher education. The survey was designed using Microsoft Forms and the interviews were conducted physically or online, based on the participants' preferences. The data collected from the interviews and survey was transcribed and coded, which allowed the connection of primary data with terms of interest [10]. This process can be classified as content analysis.

To get approval for the interview and survey, the Ethics Committee of the University of Twente was contacted. When taking part in the interview, the participants received an informed consent form, alongside an information sheet offering an overview of the project. For surveys, an opening statement has been composed that offered the necessary information for the respondents. Thus, transparency, a key factor of the research, was ensured from the beginning.

5 RESULTS

The section on results depicts the taken steps and the discovered findings throughout the research period.

5.1 Literature Review

The first research question, RQ1, is designated to help with discovering and analysing potential issues that can occur whilst using AI for the evaluation of open-ended questions. A systematic literature review was conducted to provide an exhaustive response to the question. This process started at the beginning of the research, whilst finding related works, as described in Section 3. Using the search queries from Table 1, the generated results were analysed in-depth, and the papers connected to RQ1 were selected for review.

The existing literature underlines diverse issues with AI in grading. Thus, a mind map with reoccurring concerns was created as a starting point, as seen in Figure 2.

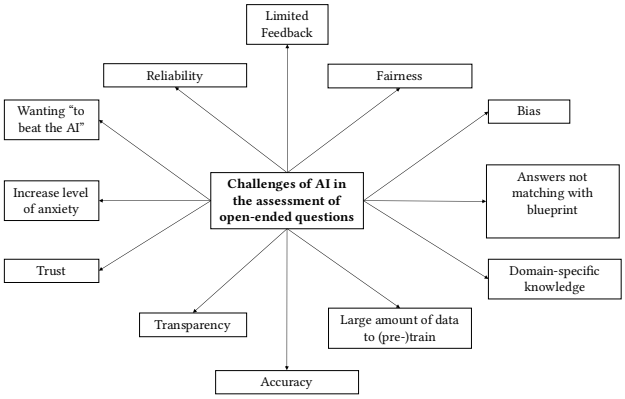

Figure 2. AI Challenges Mind Map

This section delves into the known challenges of AI in the assessment of open or short-answer questions, as illustrated in the resulting articles of the literature review.

5.1.1 Feedback

When thinking about the assessment of open-ended questions one key element is feedback. Receiving feedback on performance is "arguably the cornerstone of all learning" [19]. Looking at the literature, one visible worry is the limited amount of automated feedback received on open-ended or short answer questions. For example, detailed and personalized suggestions are not always used for automated grading tools [7, 14, 27]. When students submit their answers, the expectation would be to receive back comments that would individually target their solutions. Currently, most AI grading tools offer general remarks based on pre-set criteria. This illustrates that the used systems only reflect and quantify predetermined learning objectives [18].

5.1.2 Fairness and Bias

With the introduction of techniques, such as NLP or ML, and methods like pre-trained word embeddings [8], concerns regarding fairness and bias have been arising. In higher education, these techniques can enable the processing of substantial amounts of student data and can arrive at effective decisions in a short time [17]. Thus, a challenge encountered in using NLP or ML for a grading tool would be assuring the objectiveness and fair decision-making of the algorithms.

The issue with the used algorithms is that they are not automatically objective and unbiased [21]. Unfairness and bias are introduced through the data; therefore, they can exist in the dataset from the beginning or appear with missing/erroneous information. Hence, with ML, for example, the tool can learn new and preserve old biases [15] if it is not trained properly.

5.1.3 *Blueprints, Knowledge, and Accuracy*

From new information being used to creative answers being given for an open-ended question, a student response can cause problems in the "understanding" of an automated grading tool.

The complexity of natural language makes it "practically impossible to list all of the semantically equivalent sentences" [27] that can be used for a correct assessment of open-ended questions. Hence, one matter that affects the understanding of automated grading is the procedure in which the correct assessment for multiple possible solutions will be assured. Existing ideas in the literature discuss the training of the tool based on a large dataset of past answers [22] or the creation of subsets that will allow an alignment with the initial blueprint [18, 27]. However, there is not one existing solution that will guarantee a flawless assessment of the examined question.

Moving on from the broad aspects discussed above, the focus can be shifted towards domain-level knowledge. With an AI-supported grading tool, it is not only important to observe the variety of responses that are given by students, but also to consider the meaning of the answer and the use of the appropriate subject knowledge [9].

Another key element of automated grading is the precision and accuracy of the assessment. In case the student provided a correct answer to a question and the automated grading tool produced a wrong grade, a "false-negative assessment", the learner will waste time trying to find out what he did wrong [27]. Considering a "false-positive grade" the students will not get to know their mistakes.

5.1.4 *Trust, Transparency and Reliability*

Even if all the above-mentioned issues will be solved and the AI-supported grading tool will be a fair-decision maker and provide customized feedback, it is still up to discuss how the students will perceive this. A specified concern in the literature is the need of understanding how an AI tool is working. Students must know and understand the criteria that they are graded on, as well as acknowledge how algorithms behind the grading tool are working [18]. One paper researching the students' experiences with an automated essay scorer [24], brought to light issues regarding trust and reliability. Some students do not trust a computer to assess writing in a summative fashion. To support this claim, in the "Handbook of Automated Scoring" [30] the trend of public mistrust of automated grading is described.

Several proposals on how to alleviate these concerns are present in the literature. For example, the trust level would increase once the tool would be used for a longer period and its reliability would be visible [24]. Moreover, the students' concerns might disappear with "a greater transparency in communication and producing relevant documentation of the validity of the grading process" [30]. This approach was illustrated in a context in which the students could see what permutations of answers were considered the correct ones and then given the opportunity to ask questions to their teachers [14].

From these issues with trust, transparency and reliability, other experiences that students might have could be the need of beating the AI (e.g., finding the words that would score more points) or negative emotions caused by the inaccuracy of the tool. These types of emotions can trigger anxiety "created by someone I do not know, some algorithm which is a bit sketchy" [14].

5.2 Interviews and Survey

Once a basic understanding of the potential challenges presented in the literature was created, it was important to observe the teaching assistants' thoughts and feelings in a real-life context. Thus, the research was divided into interviews and a survey, to find answers to the second research question, RQ2. The interview and survey questions are included in Appendix A.

5.2.1 *Overview of the Study*

For the research, bachelor's and master's students who have been working as teaching assistants were contacted. In total, 12 interviews were conducted and 26 answers to the survey were received. The studies and the domains of expertise were diverse, as seen in Figure 3.

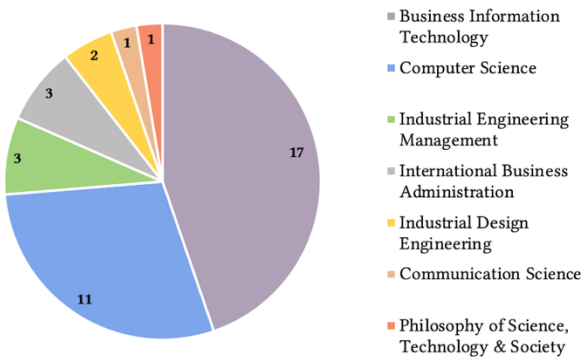

Figure 3. Research Participants

The participants were asked about their thoughts on topics such as human input for grading, automation or the implementation and use of an AI-supported tool. Numerous discussions produced meaningful insights that are described in the sections below.

5.2.2 *Human Input*

The human touch is a key factor in education. When asked about the importance of human input in grading, half of the interviewees said that it is significant to have this "to assess more than just the end result". Teachers and TAs are there to "motivate us and make the course more attractive", as discussed in one interview. The other half of the interviewees mentioned that human input is highly dependent on the type of open-ended question. For example, in the case of more applied studies, such as mathematics or programming, an open question can include code fragments, and thus can be graded based on clear criteria, not needing a human examiner. It can however be argued that better feedback can be offered by an assessor.

For the survey, this aspect is also seen as essential for the increased engagement and development of students. In the case of the survey, the respondents selected on a scale of 1-10 how important they believed the human touch to be. The majority of the answers (more than 80%) were above 7.

The benefits and issues with having a human assessor were discussed next. Table 2 shows an overview of the themes discovered during the interviews and the survey.

Table 2. Benefits and Issues with Human Assessment

Benefits	Issues
Understanding	*No clear grading criteria*
Interpretation	*Not enough knowledge*
Award partial points	*Issues with platforms used for grading*
Provide feedback	*Differences in grading of TAs*
Student – Teacher/TA Relationship	*Consistency issues*
Easily spot same mistakes	*Emotions and "Correction fatigue"*

Starting with the additional value or benefits that a person can bring compared to a tool or machine, a myriad of aspects were expressed during the interviews and survey. One of the most important ones is understanding. It is believed that a human assessor understands, interprets, and sees the nuance of the answers or the perceptions of students easier than a machine. As one of the interviewees mentioned, "You need a creative person to understand a creative answer". One respondent to the survey summed up this topic as well "The ability of humans to decipher the [handwritten] answers of students is really important. As a human, you are reading the answer 'out-loud' in your head and thereby interpreting the answer. Humans can take multiple perspectives on the answer". To illustrate this more, even the use of different vocabulary and errors in the used language are simple tasks that a human can grasp faster, as compared to an automated tool. Hence, understanding and interpreting one's answer requires intuition and expertise.

Additionally, if the student makes a mistake at the beginning of the answer and rectifies it at a later stage, a human can award partial points and provide feedback as points of improvement. This highlights the ability of humans to think outside of the box and understand context and relevance, as opposed to a machine. A respondent to the survey explained that humans "can adapt to unique answers, novelties or anomalies" and "can read the answers of students in the most charitable way", whilst a machine or tool is rigid in grading.

Furthermore, having a human grader strengthens the student-teacher/TA relationship. A student has the opportunity to discuss with a person and receive explanations as to why the answer was wrong or if it can be improved.

It is also important to note that, when having a human assessor, the issues in the question (prompt) formulation or the gaps in students' learned information can be easily observed. This was brought up by 4 of the interviewees, asking "If the same mistake has been made by all students, how can a tool deduce that the formulation of the question is bad, or all students have insufficient knowledge?".

With as many advantages that were discovered, downsides were also a theme of discussion. A serious problem that kept coming back in all interviews, was the prior knowledge of TAs. Without clear grading criteria, a part of TAs might not know the answers to the questions themselves. In the survey, this was mentioned as a grader who "might not be aware of all answers" or "with an unintentional misunderstanding of the answer". The interviews display situations of TAs creating "their version" of what the correct answer might be after reading the first few exams. This way they would grade until the end, and only then come back to the first graded questions to re-check them. The issue can become bigger in the situations in which the TAs also do not find the platform for grading "intuitive enough". In the survey, 31% of the respondents mentioned having difficulties with the used platforms.

As the grading process is very time-consuming, TAs also mentioned that one encountered problem was cross-reading of students' answers. The reason behind this is finishing grading faster. From this, differences in the grading of TAs can arise, another concept discussed during the interviews and the survey. Other potential causes of differences can be variations in understanding/ "seeing things differently", not paying more attention, and no clear guidelines. Eventually, they all create consistency issues.

Lastly, a critical aspect mentioned as a potential drawback of a human assessor were emotions or feelings, caused by factors outside the grading process. When grading, emotions, such as sadness, can be projected on the assessment "You can become judgmental if somebody upset[s] you", as voiced by an interviewee. Even more, the term "correction fatigue" was specified in the survey, and it explains that being tired can lead to not engaging properly in the grading process. There are also edge cases in which the assessor knows the student and is inclined to give them a higher grade. Other situations could be moments in which the TAs grade more students with 0 points, and they start wondering if they should change the grades. This can be applied if the exam is graded integrally, as the assessor will try to give more points for different answers, but it can also be applied when a TA is grading only one question. For example, if the first students did not score points, the assessors might reflect and examine if alternative points can be awarded, or if their grading was too strict. One interviewee explained that "Sometimes you see you give 0 points, and you are trying so hard to find where more points can be awarded".

5.2.3 Automation

Following up on the human aspect and the issues that might arise when grading, the potential of automation in the grading of open-ended questions was assessed. More than 80% of the interviewees responded that automation could help in grading, whilst in the surveys, 84% agreed with that. Some explained that automation can double-check the work of TAs and help with the repetitive tasks such that the grading process will become faster, and the students will have more time to improve their work. Others discussed feedback and allowing teachers and TAs to spend more time communicating with students instead of grading. More reasons mentioned, in the interviews, were consistency, standardization and the creation of clear guidelines which all come as an aid to the problems that TAs are currently experiencing. The survey illustrated that automation can further reduce human error and bias, can become a guiding tool for the assessors, "flagging completely wrong or correct answers, similar to a pre-process before a TA is checking" and it will be less prone to human fatigue.

Looking at why automation cannot help, the remaining respondents of the interviews, argued that if automation can reach a high level of accuracy, the jobs of TAs will disappear. As for the survey, respondents claimed that "No machine can be trained to judge everyone's unique thoughts", as those might be "all over the place" and "not structured enough".

5.2.4 AI-Supported Tool

Moving from automation, the implementation of an AI-supported tool was discussed. The interviewees were encouraged to think of their experience with intelligent tools such as CodeGrade, Remindo or SpeedGrader (platforms that aid assessors in the grading of student's exams or exercises).

The topic of an AI-supported grading tool brought to light some potential advantages and challenges that might be encountered by teachers, TAs and students.

Looking at benefits, efficiency, and accuracy in terms of quality of the grading, feedback received, and time were mentioned. By far, the increase in the speed of grading was one of the most mentioned topics, in both the interviews and surveys. The interviewees explained that having such a tool can "take out repetitive tasks" or "saves time, with clear guidelines". In an expert interview [28] the participants also agreed that grading is labour-intensive. The survey highlighted that faster grading will have positive effects on both the teachers and the students. One of the respondents stated that "The tool decreases the grading time of the examiners", whilst another mentioned, "Students get answers faster".

Furthermore, it is important to note that the tool is believed to support TAs in several cases. For example, when different

assistants will be grading, the tool will make the process fairer, considering that the emotions of humans will be excluded, or favouritism will be left out, as one of the interviewees mentioned "AI is better at being objective". Additionally, in the cases in which new TAs are afraid to communicate or ask questions to more experienced assistants, "such a tool can become an aid". Hence, the tool is considered less subjective than a human, thus making the grading process fairer, consistent, and excluding certain biases.

Challenges and drawbacks of such a tool were also debated. A general direction of the interviewees was that in the beginning a lot of time and investment will be necessary to make sure that the tool is matching the teacher and the subject. This will later be followed by extensive testing, which is also considered time-consuming. If the tool will not be properly configured, the AI can represent "another person that you will have to check upon" as told by an interviewee, thus increasing the time spent grading. In the survey, it was also explained that additional reviews for the tool can be extremely labour-intensive.

The idea that the AI should only be used as a support tool stood out in 11/12 interviews, this way the human side will not be removed. The remaining participant wanted a fully autonomous AI tool. As for the survey, respondents mentioned that a hybrid way of grading is more beneficial than a standalone tool. Moreover, another important aspect discussed was the "laziness" of TAs. An interviewee revealed that "TAs are likely to take the grading done by the AI for granted". In the survey, the complaints about the laziness of assistants were also mentioned "TAs will become lazy, and this can cause inaccurate grading", however, this was only mentioned twice.

Other discussed concerns were the exploitation of the tool by the students, having the AI become biased without people noticing (wrong answers can become correct ones if the system learns inaccurately) or the fact that current platforms are not intuitive enough (mistakes can happen with deletion of files, people cannot work at the same time etc.). To go more in-depth about bias, the survey revealed other potential types of biases such as preference for technical answers, short or long answers, or grammar over the content.

Possible ways to overcome the challenges were also talked about. The main aspect would be to train the model extensively and to focus on multiple languages, characters, and syntax such that the algorithm can run smoothly. This would mean having a well-structured training dataset and performing incremental changes to continuously improve the AI. The user-friendliness, stability, and intuitiveness must be considered prior to the use of the AI tool. For the concern of laziness, creative solutions were identified. One of them was to use random picking and leave one of the questions ungraded by the AI. This way, it can become visible if a TA has rigorously checked the questions or just blindly accepted the work of the AI. Another solution was to have the AI as a way to examine the TAs' work, instead of having the TAs inspect the AI. However, both approaches are unlikely to happen, as they might negatively impact the teaching assistants.

The last discussion point on the AI-supported tool was its implementation for different subjects. It was revealed that it is not the subject that matters, but the type of question that is being asked. There was a consensus that knowledge questions, where definitions of terms and relationships are required, or algorithmic ones for mathematic formulas and programming exercises are asked, would benefit more from such a tool, than higher-order questions. This happens because they are easier to standardize "look at keywords", and the "algorithm can know all the steps". For open questions that entail more of a personal opinion, it is believed that a student's answer will always be different from the one of the teachers or from one of the other students, thus harder to use an AI. Three of the interviewees mentioned that they would keep an AI tool in this situation only for "plagiarism checks or structure guidelines".

Table 3 illustrates the beliefs on the most important aspects mentioned in the interviews and the survey.

Table 3. Mentions of Benefits and Drawbacks with an AI-supported Tool

Mentions		Interviews	Survey
Main Benefits and Issues of an AI-Tool	*(+) Faster grading*	9/12 (75%)	24/26 (92%)
	(+) Support	11/12 (92%)	10/26 (38%)
	(-) A lot of time, Computational power, Initial investment	8/12 (66%)	17/26 (65%)
	(-) Influence of AI over a person (emotions, laziness, "correction fatigue")	8/12 (66%)	2/26 (8%)
Other Mentions	*(+) Quality of feedback, Consistency, Reduced bias of humans, Better at being objective*		
	(-) Exploitation, Bias of AI without noticing, Platform/AI not intuitive and user friendly		

5.2.5 *Attitude*

With an AI-supported grading tool, it is important to observe what might be the attitude of TAs and students.

Figure 4 offers an overview of the discovered attitudes and beliefs.

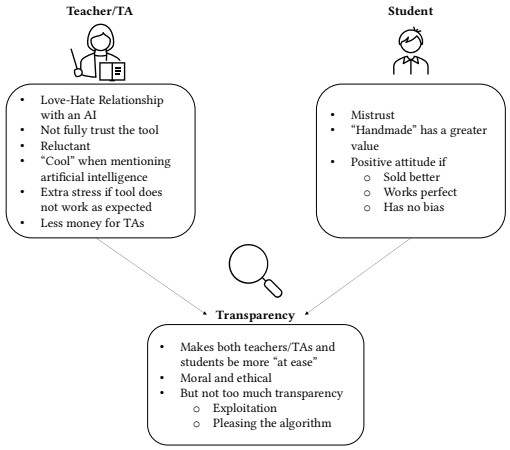

Figure 4. Attitude of Teachers/TAs and Students

On one hand, for TAs, the tool will not be fully trusted at the beginning, but they might gradually adapt to it. On the other, it can be perceived with curiosity and excitement, thinking that it is something "cool" as one respondent to the survey and two interviewees mentioned. The use of the tool can also cause extra work and stress for TAs if it is not properly implemented, and no explanation has been given on how to use it to its full capacity.

Furthermore, when presented with such a tool, the assistants can be reluctant and ironically say "Oh, great, another tool", as mentioned by one of the interviewees. Another discussion during an interview illustrated that "people are not big fans of changes" and this supports the above-mentioned perspective. It is believed that a "love-hate relationship will appear" as TAs will work faster, but they will receive less money. The issue of payment was mentioned in more than 50% of the interviews. However, only 4 survey responses explained that when the work will be reduced, payment issues will start appearing and the authority of the TAs will diminish.

From the student's perspective, every interview discussed the fact that they will not trust an AI to fully grade them. Trust was

a voiced concern in the responses of the survey as well. One of the interviewees mentioned that "From the psychological approach, they would like to know that a human being is in front of [a] screen who reads the entire answer, understands it, and gives feedback". It was also explained that a human-to-human relationship means trust and "As people appreciate and pay more for handmade objects [compared to mass production], this way they will [also] appreciate the human input more". Hence, the general observation is that an AI-supported tool will be perceived negatively, especially if the grade will be bad, if a student will have to work more to defend their grade, or if no individualized feedback will be given.

This attitude can also escalate with mistrust and not knowing who graded the test, a TA, or an AI. The students might perceive this in a more positive manner if the tool will be "sold better than just an AI tool", as one of the interviewees mentioned, or if they will know that "the tool is working 100% and the grades come faster and have no bias".

The issue of trust might be solved with the help of a higher degree of transparency. This way the acceptance and reliability in education will be influenced positively and bring the students "a bit more at ease". Increasing transparency will also bring a plus to the moral standpoint. However, a few of the interviewees strongly claimed that if a student is sceptical or fails, no matter the degree of transparency, the tool will still be perceived in a negative manner. Furthermore, too much transparency can lead to exploitation "people will know what to expect" and they will start "pleasing the algorithm" instead of focusing on understanding the topic.

5.2.6 *Alternatives*

The last question made the respondents think divergently. They were asked if they could name alternatives to AI that could support teachers in the grading of open-ended questions. Prerequisites for an AI-supported tool could thus be observed, as most of the interviewed TAs mentioned having available clearer and more explicit grading criteria or rubrics. They felt like guidelines and examples are missing most of the time. For example, explicit subtasks can be created for them to know "what and where to look" at and observe the desired situation or context. Another viable solution would be having review sessions with both teachers and TAs to increase collaboration and have their questions explained in due time.

One alternative brought up by an interviewee was to create more competition between universities. It was explained that in a company context there exists competition in order to make them grow (e.g., each company wants better employees for the best/most optimal results). By creating a competition between universities, the teachers and TAs will also be more determined to grade and enhance students' learning journey. In the context of grading, an AI-supported tool will provide grades faster and in a more accurate manner, thus positively influencing the experience of students and potentially "making them more competitive". Nonetheless, with competition, a matter in question would be how to assure that grades are offered fairly. Here, a potential use of an AI tool can help to standardize the process and be objective.

5.3 Prerequisites of an AI-supported Tool

After analysing aspects related to AI, found in the literature, as well as, during interviews and a survey, connections and patterns were created as a potential response to the main research question. Figure 5 offers an overview of suggested prerequisites such that AI can help in the grading of open-ended questions. Each aspect is described in the subsequent sections.

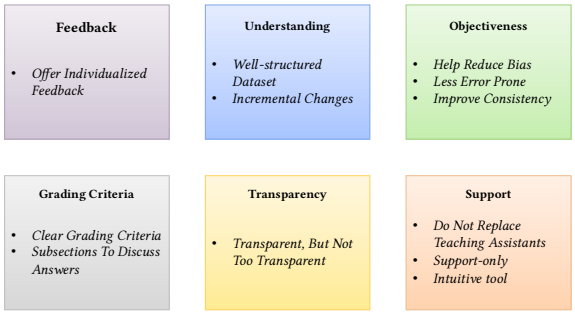

Figure 5. Prerequisites of an AI-supported Tool

5.3.1 *Common Themes*

The most discussed aspect in the literature was feedback. Currently, existing AI tools do not offer automated individualized feedback when grading a student's answer. This topic was not a focal point in the interviews and survey, as only a few respondents explicitly mentioned it. However, having the ability to offer feedback should be a prerequisite of AI in the grading of open-ended questions. With this capacity, the assessors can spend less time writing comments for each student.

Understanding was the most specified aspect of the survey and interviews. With understanding, everything from the use of creative answers to ambiguity and errors has been referred to. Compared to a machine a human assessor can have a better understanding of a student's answer. The issue is also discussed in the literature, and it can lead to substantial time and investment for the proper usage of an intelligent tool. To make sure that an AI will understand a response to a question, the dataset should be well structured and contain numerous past answers that the algorithm can learn upon. The model should also be trained extensively, incremental changes should be made, and domain-specific knowledge should be used, to assure the best results of the algorithm. With such a dataset, the tool is believed to improve the precision, accuracy, and efficiency of grading.

The issues of fairness, bias and consistency must also be considered when thinking of desirable prerequisites for an AI tool. Discussed in the literature, in the survey and in the interviews, making sure that the AI can be as close to an "objective judge" as possible is essential. This way, the tool can help create a consistent grading, make the grading less prone to human error and reduce bias. However, prior to having the aid from the tool, the teacher must make sure that for the open-ended questions there are clear grading criteria, even split into subsections, that do not leave much room for bias or interpretation.

Furthermore, the impact on both the teaching assistants and the students has been analysed. The literature had limited information on what the attitudes and beliefs of students and TAs might be. Some ideas illustrated were that everyone must understand how the AI is working: the criteria of grading or even the algorithm. This perception of high transparency, in the literature, is believed to ease concerns of students regarding the tool and make them trust it more. Although, transparency is illustrated as "an ideal" in the literature, during the interviews and surveys this was debated. With too much transparency, the tool can be exploited, and the students will focus on "beating the AI" instead of on understanding the study material. Moreover, it has been considered that even with a higher degree of transparency the students who are sceptical are unlikely to change their view, as opposed to the beliefs in the literature.

Therefore, as one of the interviewees mentioned "Transparency should be offered, but not too much".

5.3.2 *Unique Prerequisite*

New insights appeared during the interviews and the survey, where the human has been seen as central. Concerns regarding people's well-being have been reflected upon.

The fear that TA jobs might disappear, or that payment will be lower if an AI-supported tool would work as close to perfect was repeatedly mentioned. On this note, one important prerequisite would be for the tool to only support TAs. Instead of removing their job and authority, they can be helped by this. For example, the AI can flag certain parts of an answer and allow the TAs to efficiently check them.

Furthermore, the tool should be intuitive and well-explained to the TAs. In a lot of the interviews, issues with the grading platforms were described, hence it is important to have a working tool that can offer reliability and support for the graders. For example, some of the current problems were no structured comment sections (e.g., use comments per sub-sections, not one general remark), stability issues (numerous files being deleted and having to re-grade everything) or the incapacity to grade the exact student assessment with another TA at the same time.

6 CONCLUSION

To bring all the findings together, the research discovered the challenges that can arise with using AI for automated grading. This was done by investigating the issues present in the literature and by conducting interviews and a survey to observe the opinions of TAs on the topic. Based on the discoveries, an overview of prerequisites that should be considered throughout the automated grading process was created.

It is important to note that, at the current moment multiple challenges exist with the implementation of AI in the grading of short answer questions. No such thing as a "perfect implementation" is available. However, for every challenge, it is believed that approaches to overcome it exist. In order to make an AI-supported tool as accurate as possible, incremental changes must always be made. By doing this, the technology acceptance will also increase, and people will feel more "at ease" or comfortable with using such a tool.

Throughout the research, it was interesting to observe that no clear difference in the perception and opinions of different study programmes was recognized. Both technical (e.g., Computer Science, Industrial Engineering and Management) and non-technical studies (e.g., Communication Science, Philosophy of Science, Technology & Society) have participated in the research. One could have expected that technical studies would have a more positive attitude towards such a tool, since it might be easier to analyse more structured responses than creative ones, present more frequently in non-technical studies.

During the research, it was discerned that human input is valued more than artificial intelligence. With the use of AI, the majority of the respondents indicated that it should only be a support tool for assessors. This way, the jobs of TAs will not disappear, and fair payment can still be made. In terms of what support means, each interviewee or respondent imagined it in a unique way, for example: help with flagging issues, such that the assessor can have an easier time navigating the answer or having an AI check the work of the assessors to assure the fairness of the grades. Thus, this makes it visible that the TAs should remain involved in the grading process considering that a tool can have various purposes.

Moving to the student's perspectives, their attitudes will vary depending on numerous factors (e.g., the grade they will be receiving). It can be anything from positive to negative and it can alter depending on the technology acceptance and transparency levels. With transparency, it is essential to assure that it cannot generate exploitation.

Overall, the direction in which all the responses were going was that automation, or AI, is the future. Presently, the best approach to integrating AI in education is in a hybrid manner. As mentioned in one of the surveys, "This could be a way for TAs to verify their grading. But it could also be the other way around, where TAs grade questions to check if their assessment matches with the ones of the AI".

7 DISCUSSION

7.1 Limitations

Throughout the study, more than half of the participants were studying Business Information Technology (BIT) and Computer Science (CS), at the University of Twente. This was influenced by the background of the researcher which is BIT. Contacting people from within the same faculty (in this case Electrical Engineering, Mathematics and Computer Science), with whom one interacted before, has proven to be an easier task in terms of responses and engagement. The number of participants from the two studies, 17, respective 11, is greater than the ones of the other populations, followed by 3, 2 or 1. Furthermore, a slight bias can also be seen towards more technical studies.

Hence, it would have been enthralling to discuss with more studies from the university, or with more people from within one study. This way a more accurate understanding of the topic could have been gained. Due to the limited timeframe of the research, this was hard to accomplish as unknown people are not always inclined for participating in such research.

7.2 Future Work

During the research, the literature, and the opinions of the respondents to both the interview and the survey brought to light avenues that could be explored in the future.

One attractive opportunity that could be investigated is the implementation of an AI-supported tool at a university-wide level, in the Netherlands. For example, how would such a tool, if standardized and used by numerous universities in the country, impact the assessment of students. Could it help with generating impartial results? Or can it improve the overall performance of the students?

Another path of inquiry is the change in the answers to the questions once such a tool would be implemented. The use of the AI-supported tool in various contexts, for example with students knowing or not that they are graded with the help of AI, or the tool executing diverse support tasks, might lead to different results. Furthermore, the students can be asked if they intend to use the tool further and if they perceived it in a positive manner. With such an approach, incremental changes can be made to assure the correct implementation and use of the tool.

Lastly, delving into the use of an AI-supported tool for more than textual open questions could be insightful. In the data collection phase of the research, it was noticed that different studies have their preferred way of assessing students. For example, in the case of Industrial Design Engineering, a course requires students to draw certain elements. If drawing for an exam could be considered an open-ended question, the potential use of AI in such an area might also be an aid to the assessors.

REFERENCES

[1] Aldea, A.I., Haller, S.M. and Luttikhuis, M.G. 2020. Towards grading automation of open questions using machine learning. *SEFI 48th Annual Conference Engaging Engineering Education, Proceedings*. (2020), 573–582.

[2] Bernius, J.P., Kovaleva, A., Krusche, S. and Bruegge, B. 2020. Towards the Automation of Grading Textual Student Submissions to Open-ended Questions. *ACM International Conference Proceeding Series*. (2020), 61–70. DOI:https://doi.org/10.1145/3396802.3396805.

[3] Burrows, S., Gurevych, I. and Stein, B. 2015. *The eras and trends of automatic short answer grading*.

[4] Busetto, L., Wick, W. and Gumbinger, C. 2020. How to use and assess qualitative research methods. *Neurological Research and Practice*. 2, 1 (2020). DOI:https://doi.org/10.1186/s42466-020-00059-z.

[5] Coulthard, G.J. 2016. A descriptive case study: Investigating the implementation of web based, automated grading and tutorial software in a freshman computer literacy course. *ProQuest Dissertations and Theses*. (2016), 208.

[6] Dadi, R., Pasha, S.N., Sallauddin, M., Sidhardha, C. and Harshavardhan, A. 2020. An overview of an automated essay grading systems on content and non content based. *IOP Conference Series: Materials Science and Engineering*. 981, 2 (2020). DOI:https://doi.org/10.1088/1757-899X/981/2/022016.

[7] Daradoumis, T., Marquès Puig, J.M., Arguedas, M. and Calvet Liñan, L. 2019. Analyzing students' perceptions to improve the design of an automated assessment tool in online distributed programming. *Computers and Education*. 128, September 2018 (2019), 159–170. DOI:https://doi.org/10.1016/j.compedu.2018.09.021.

[8] Erickson, J.A., Botelho, A., Peng, Z. and ... 2021. Is it Fair? Automated Open Response Grading. *… Data Mining*. (2021), 2–7.

[9] Erickson, J.A., Botelho, A.F., McAteer, S., Varatharaj, A. and Heffernan, N.T. 2020. The automated grading of student open responses in mathematics. *ACM International Conference Proceeding Series*. (2020), 615–624. DOI:https://doi.org/10.1145/3375462.3375523.

[10] Fossey, E., Harvey, C., Mcdermott, F. and Davidson, L. 2002. Understanding and evaluating qualitative research. *Australian and New Zealand Journal of Psychiatry*. 36, (2002), 717–732.

[11] Galassi, A. and Vittorini, P. 2021. Automated feedback to students in data science assignments: Improved implementation and results. *ACM International Conference Proceeding Series*. (2021). DOI:https://doi.org/10.1145/3464385.3464387.

[12] Galhardi, L.B. and Brancher, J.D. 2018. Machine Learning Approach for Automatic Short Answer Grading: A Systematic Review. *Lecture Notes in Computer Science*. 380–391.

[13] Ghosh, S. 2010. Online Automated Essay Grading System as a Web Based Learning (WBL) Tool in Engineering Education. *Web-Based Engineering Education*. IGI Global. 53–62.

[14] Huey, S., Tan, S., Rajalingam, P., Chia, A. and Chew, Y. 2022. Enabling open-ended questions in team-based learning using automated marking : Impact on student achievement , learning and engagement. March (2022), 1–13. DOI:https://doi.org/10.1111/jcal.12680.

[15] Kleinberg, J., Mullainathan, S. and Raghavan, M. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. (Sep. 2016).

[16] Kuleto, V., Ilić, M., Dumangiu, M., Ranković, M., Martins, O.M.D., Păun, D. and Mihoreanu, L. 2021. Exploring opportunities and challenges of artificial intelligence and machine learning in higher education institutions. *Sustainability (Switzerland)*. 13, 18 (2021), 1–16. DOI:https://doi.org/10.3390/su131810424.

[17] Marcinkowski, F., Kieslich, K., Starke, C. and Lünich, M. 2020. Implications of AI (un-)fairness in higher education admissions. (2020), 122–130. DOI:https://doi.org/10.1145/3351095.3372867.

[18] Mirmotahari, O., Berg, Y., Gjessing, S., Fremstad, E. and Damsa, C. 2019. A case-study of automated feedback assessment. *IEEE Global Engineering Education Conference, EDUCON*. April-2019, c (2019), 1190–1197. DOI:https://doi.org/10.1109/EDUCON.2019.8725249.

[19] Orrell, J. 2006. Feedback on learning achievement: Rhetoric and reality. *Teaching in Higher Education*. 11, 4 (2006), 441–456. DOI:https://doi.org/10.1080/13562510600874235.

[20] Pérez-Marín, D., Pascual-Nieto, I. and Rodríguez, P. 2009. Computer-assisted assessment of free-text answers. *Knowledge Engineering Review*. 24, 4 (2009), 353–374. DOI:https://doi.org/10.1017/S026988890999018X.

[21] Pessach, D. and Shmueli, E. 2023. A Review on Fairness in Machine Learning. *ACM Computing Surveys*. 55, 3 (2023), 1–44. DOI:https://doi.org/10.1145/3494672.

[22] Ramesh, D. and Sanampudi, S.K. 2022. *An automated essay scoring systems: a systematic literature review*. Springer Netherlands.

[23] Roy, S., Narahari, Y. and Deshmukh, O.D. 2015. A perspective on computer assisted assessment techniques for short free-text answers. *Communications in Computer and Information Science*. 571, (2015), 96–109. DOI:https://doi.org/10.1007/978-3-319-27704-2_10.

[24] Scharber, C., Dexter, S. and Riedel, E. 2008. Students' experiences with an automated essay scorer. *Journal of Technology, Learning, and Assessment*. 7, 1 (2008), 1–44.

[25] Schilling, N. 2013. Surveys and interviews. *Research Methods in Linguistics*. Cambridge University Press. 96–115.

[26] Smith, M.A. and Karpicke, J.D. 2014. Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*. 22, 7 (2014), 784–802. DOI:https://doi.org/10.1080/09658211.2013.831454.

[27] Sychev, O., Anikin, A. and Prokudin, A. 2020. Automatic grading and hinting in open-ended text questions. *Cognitive Systems Research*. 59, (2020), 264–272. DOI:https://doi.org/10.1016/j.cogsys.2019.09.025.

[28] TA1 2022. *Expert interview with teaching assistant from IBA at University of Twente*.

[29] Wang, H.C., Chang, C.Y. and Li, T.Y. 2008. Assessing creative problem-solving with automated text grading. *Computers and Education*. 51, 4 (2008), 1450–1466. DOI:https://doi.org/10.1016/j.compedu.2008.01.006.

[30] Yan, D., Rupp, A.A. and Foltz, P.W. 2020. *Handbook of Automated Scoring*. Chapman and Hall/CRC.

APPENDICES

A APPENDIX A – INTERVIEW & SURVEY QUESTIONS

- What is your **study**?
- How important do you think **human input** is for grading open-ended questions?
    - Why do you think that is important?
    - Can you name a few benefits of having a teacher/student grading an open-ended question instead of a machine/tool?
    - Are there any issues that can arise when assessing open-ended questions?
      Note: Think from your experience as a teaching assistant. If there are no cases in which an issue arose, you can imagine what might happen.
- Do you believe that **automation** can help with the assessment of open-ended questions?
    - Why/Why not?
- Context: Let's suppose that an **AI-supported grading tool** is being implemented at the University of Twente. The purpose of this tool will be to support examiners in the grading of open-ended questions.
    - What might be a benefit of such a tool?
    - And what do you think the drawbacks/challenges would be with the AI-supported grading tool? Note: You can name anything you can imagine; As a starting point you can think of your experience with tools such as CodeGrade, SpeedGrader or Remindo.
    - Any way of overcoming the challenges?
    - Do you believe that having such a tool would be easier to implement and use for particular subjects? Why? Note: Consider various areas of study in which such a tool might be applied. (e.g., mathematics, ethics/philosophy etc.)
- From your perspective what would be the **general attitude** of teaching assistants when this tool would be implemented?
    - Do you think that students who did not work as teaching assistants might perceive this differently? Why/Why not?
    - Would having higher transparency, for example, a clear explanation of the algorithms, help the TAs or the students perceive the AI-supported tool in a different way?
- Do you think there are better **alternatives** than AI to support the teachers/TAs in the grading of open-ended questions?