Explainability in Automatic Short Answer Grading

Tim Schlippe, Quintus Stierstorfer, Maurice ten Koppel, Paul Libbrecht

IU International University of Applied Sciences
tim.schlippe@iu.org

Abstract. Massive open online courses and other online study opportunities are providing easier access to education for more and more people around the world. To cope with the large number of exams to be assessed in these courses, AI-driven automatic short answer grading can recommend teaching staff to assign points when evaluating free text answers, leading to faster and fairer grading. But what would be the best way to work with the AI? In this paper, we investigate and evaluate different methods for explainability in automatic short answer grading. Our survey of over 70 professors, lecturers and teachers with grading experience showed that displaying the predicted points together with matches between student answer and model answer is rated better than the other tested explainable AI (XAI) methods in the aspects *trust*, *informative content*, *speed*, *consistency and fairness*, *fun*, *comprehensibility*, *applicability*, *use in exam preparation*, and *in general*.

Keywords: explainability, explainable AI, XAI, automatic short answer grading, AI in education.

1 Introduction

Access to education is one of people's most important assets and ensuring inclusive and equitable quality education is goal 4 of United Nation's Sustainable Development Goals [1]. Distance learning in particular can create education in areas where there are no educational institutions or in times of a pandemic. There are more and more offers for distance learning worldwide and challenges like the physical absence of the teacher and the classmates or the lack of motivation of the students are countered with technical solutions like videoconferencing systems [2] and gamification of learning [3]. The research area "AI in Education" addresses the application and evaluation of Artificial Intelligence (AI) methods in the context of education and training [4, 5, 6]. One of the main focuses of this research is to analyze and improve teaching and learning processes. Many educational institutions—public and private—already conduct their courses and examinations online. This means that student examinations and their assessments are already available in digital, machine-readable form, offering a wide range of analysis options. An exam often consists of multiple choice and free text questions. Multiple choice questions can be made so that answers are unambiguous and have been easy to evaluate by machine for many years. However, the evaluation of free text answers—

i.e., assigning quantitative feedback on the correctness of the student response in the form of a score possible within a certain point range—still required tedious manual work by the graders since it was a greater challenge for automatic systems. Fortunately, automatic short answer grading (ASAG) is improving and in some cases has already reached the point where teaching staff could use it for fairer and faster grading [7]. Since the results in terms of scores are not yet perfect and graders want to understand ASAG systems' decisions, the question arises: What is the best way to make the decisions of ASAG systems explainable to human graders? The desire for explainability is also demonstrated by the feedback from 71 professors, lecturers, and teachers, as shown in Figure 1. A clear majority of participants strongly agrees that it is important for them to understand how an AI reaches its expected scoring (4.42 on average) and their confidence in an AI grading support increases when it explains itself (4.18 on average).

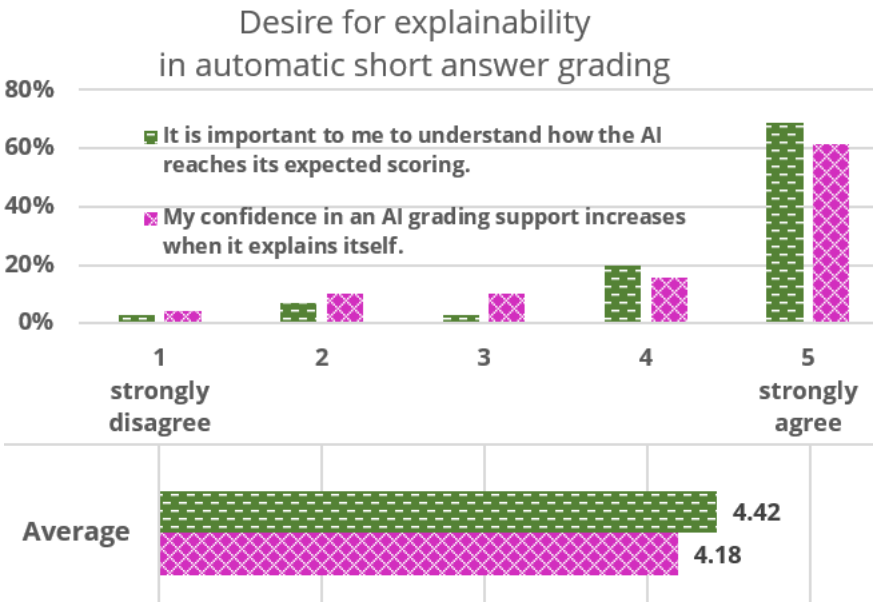

Fig. 1. Desire for explainability in automatic short answer grading.

The field of explainable AI (XAI) aims to provide solutions for the need of transparency. XAI can be described as research direction that aims to create human interpretable AI [8]. In this paper, we investigate and evaluate different methods for explainability in ASAG. Thus, we provide insight into the perceived usefulness of different XAI methods for ASAG. For the evaluation of the XAI methods we asked over 70 professors, lecturers, and teachers to rate different aspects.

In the next section, we will present the latest approaches of other researchers for ASAG and explainability. Section 3 will demonstrate our investigated methods for explainability in ASAG. Section 4 will describe the experimental setup for our user

study. The study and the results are outlined in Section 5. We will conclude our work and suggest further steps in Section 6.

2 Related Work

In this section we will present related work in the areas of explainability and ASAG.

2.1 Explainability

Before the deep learning era, data scientists crafted predictive models by manually inspecting data and constructing models based on the insights. With deep learning, the best practice is now to let the algorithm figure out itself which parts of the data are useful [9]. Typically, modern deep learning models are created with hundreds of features using gigabyte-sized data sets. Verification of these models is usually done by calculating accuracy: The goal is that a model makes as few errors as possible ignoring the reasons why a prediction is made. However, experts are now beginning to understand that this metric alone is not enough [10]. We are beginning to ask questions such as: Is a model robust to small variations in data? How does the model behave when atypical data is input? Is a model safe to use or could data be extracted? Does the model respect privacy of individuals? Is it fair and non-discriminatory?

The field of XAI aims to provide solutions for these transparency needs. Although the field lacks a clear definition, it can be best described as a research direction that aims to create human interpretable or explainable AI [1]. The human interpretation part of the explanation is critical: If it cannot be understood or applied by a human, it becomes meaningless. The field of XAI is therefore considered to be a multi-disciplinary field in which data scientists, AI engineers, human scientists, and human-computer interaction specialists work together to create technical sound and human interpretable explanations. XAI has a similarly long history like the field of AI itself. However, it first started to boom after the deep learning revolution in 2012 [11]. Good overviews of the field are provided by [8,12,13,14,15].

On the most basic level, two directions can be identified within this field: i) Methods that aims to make models intrinsically interpretable and ii) methods that aim to provide transparency for black box models (post-hoc methods). The latter is the vast dominating approach in the field. Black box models are models that "are created directly from data by an algorithm, meaning that humans, even those who design them, cannot understand how variables are being combined to make predictions" [16]. To deal with those black box models, ongoing research is the creation of additional predictive models that can meet the accuracy of black box models while being intrinsically interpretable.

Since the state-of-the-art ASAG models are mostly based on transformer models, they are also black box models. Therefore, our XAI methods are not directly based on an interpretation of the actual existing ASAG model but use separate models for explainability.

2.2 Automatic Short Answer Grading

The field of ASAG is becoming more relevant since many educational institutions—public and private—already conduct their courses and examinations online [7,17]. A good overview of approaches in ASAG before the deep learning era is given in [18]. [19] and [17] investigate and compare state-of-the-art deep learning techniques for ASAG. [19] demonstrate that systems based on BERT performed best for English and German and that their multilingual RoBERTa model [20] shows a stronger generalization across languages on English and German. [7] extended ASAG to 26 languages and use the smaller M-BERT [21] model to conduct a larger study concerning the cross-lingual transfer. With Mean Absolute Errors between 0.41 and 0.72 points out of 5 points they report that their best models have even less discrepancy than 2 graders, which is 0.75 points. This shows that the performances are now good enough to be used as a support for scoring answers to open exam questions—provided that the prediction of AI is presented in a good way to the graders.

A first work towards explainability in ASAG is described by [22]. However, their focus is a comparison of different corpora and feature attribution techniques to "attribute" words in the student answer, which should then make the decision of the ASAG system more explainable. While the focus of [22] is on the technical implementation of their attribution values with transformer-based models, our work is more general as it does not depend on specific models, evaluates the perception of a broader range of XAI methods and has a strong focus on feedback of the teaching staff through detailed evaluation of 9 aspects as described in Section 4 and 5. We can imagine that our results can be used complementary to the results of [22].

3 Methods for Explainability in Automatic Short Answer Grading

The goal of ASAG is to enable (semi) automatic, fair, and consistent grading at a high productivity. We explored if and how explanations might be beneficial to this use case. Figure 2 visualizes with an example the pipeline of our systems for AI-driven grading support. The system consists of 1 ASAG model for point prediction and 1 model for explainability. The ASAG model always processes the student answer and the model answer[1] for the prediction. Depending on the XAI model, only the student answer or the student answer plus the model answer is input. In our experiments, we exchanged the XAI method so that its output—here *Matching Positions*—is different for each method.

[1] also called *sample answer* or *sample response* in literature

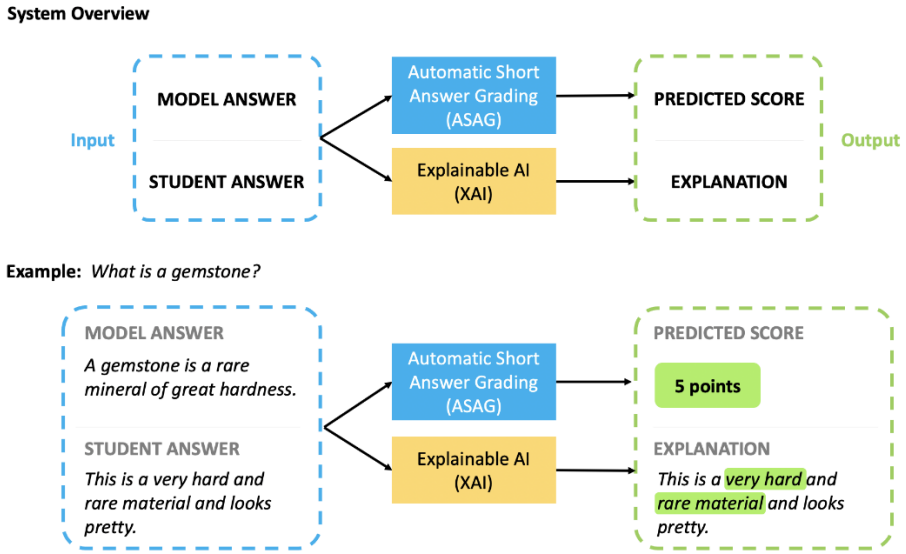

Fig. 2. Pipeline for grading with point prediction and explainability.

In our investigation we examined common XAI method classes described in the literature which could be adapted to the ASAG task—even if they have only been used for Computer Vision and not yet for Natural Language Processing (NLP). Table 1 demonstrates 3 prevalent common XAI method classes which are appropriate for this task and describes their characteristics.

Table 1. XAI method classes suitable for the generation of XAI methods specific for ASAG.

XAI method class	Description
confidence score	Certainty of a model's prediction is made interpretable and inspectable [23].
word highlighting	Words are color marked to indicate their relevance towards the classification [24].
concept activation	High level human concepts are used to explain a classification [25].

From the remaining XAI method classes, we created more specific methods useful for the ASAG task. Table 2 lists our 5 created XAI methods along with the method classes on which their creation is based. Only *Predicted Points* is not based on a method class, since only the number of points is displayed.

Table 2. AIX methods specific for ASAG.

XAI method for ASAG	XAI method class
Predicted Points	—
Predicted Points with Confidence Scores	*confidence score*
Predicted Points with Confidence Scores and Similar Answers	*confidence score*
Predicted Points with Relevance of Words in the Answer	*word highlighting*
Predicted Points with Matching Positions	*concept activation*

In the next sections, our developed ASAG-specific XAI methods are described in more detail.

3.1 Predicted Points

As illustrated in Figure 3, displaying only the predicted points is the simplest of the evaluated methods (*Predicted Points*). No additional XAI model is used.

Fig. 3. XAI Method: Predicted Points.

3.2 Predicted Points with Confidence Scores

As demonstrated in Figure 4 (*Predicted Points with Confidence Scores*), interpretable confidence scores put the predicted score in context of past performance. This confidence score is computed by considering similar past cases where the AI and human collaborated, and hence the system has feedback on the accuracy of its predictions. The interpretable confidence score is essentially a single percentage between 0% and 100%, with additional information on how many similar answers it is computed.

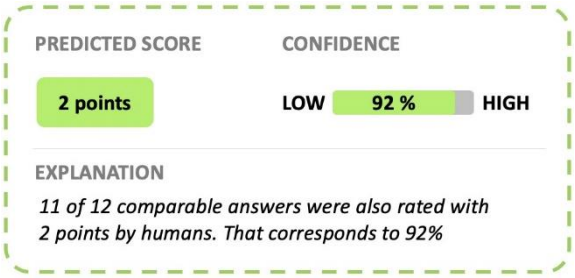

Fig. 4. XAI Method: Predicted Points with Confidence Scores.

3.3 Predicted Points with Confidence Scores and Similar Answers

Interpretable confidence scores can be extended by providing examples of answers that are similar as shown in Figure 5 (*Predicted Points with Confidence Scores and Similar Answers*). Hence, the graders do not only get one confidence score but also see examples of answers that were rated equally. This concept with comparable scored answers could also help new graders get into the grading process more quickly as well as help students relate their answer to other answers and deduce why their own answer is incorrect or correct.

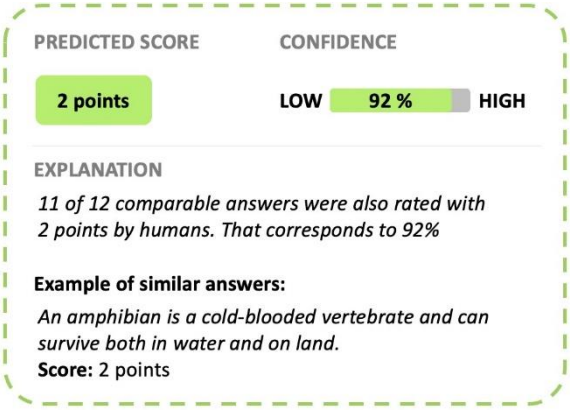

Fig. 5. XAI Method: Predicted Points with Confidence Scores and Similar Answers.

3.4 Predicted Points with Relevance of Words in the Answer

The standard method for explainability in the context of NLP is word highlighting [14, 24]. Typically, every word is marked to indicate its relevance for the prediction. However, this can be confusing and hard to interpret. Instead in this method a threshold of relevance makes sure that only a subset of the most important words is marked as indicated in Figure 6 (*Predicted Points with Relevance of Words in the Answer*). The idea is that this leads to a more efficient interpretation of information.

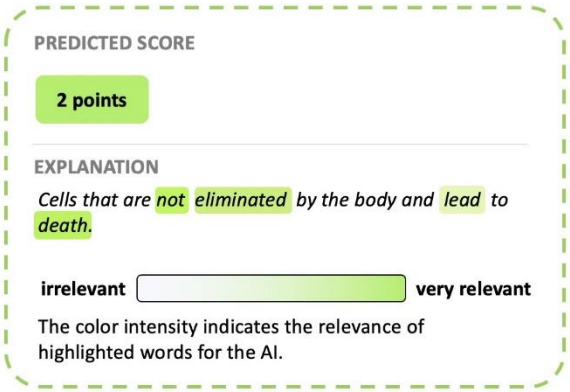

Fig. 6. XAI Method: Predicted Points with Relevance of Words in the Answer.

3.5 Predicted Points with Matching Positions

The method of concept attribution builds on the idea that a good explanation relates to understandable human concepts [25]. In its application for computer vision, not pixels themselves are highlighted, but instead an image is analyzed on the availability of a human understandable concept, e.g., the presence of medical condition. This method has not yet been demonstrated in NLP.

To transfer this idea to ASAG, we propose that the correct parts of the model answer are highlighted within the student answer (*Predicted Points with Matching Positions*) as shown in Figure 7. In essence, the model answer is already a human understandable concept which an explanation should ideally relate to. We believe that this method leads to an efficient interpretation of information which should be useful to both the grader and the student.

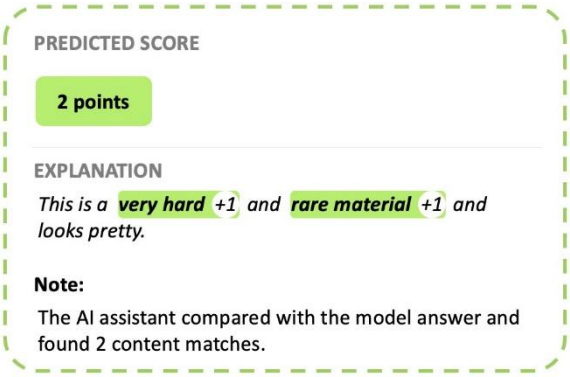

Fig. 7. XAI Method: Predicted Points with Matching Positions.

4 Experimental Setup

As mentioned before, we initially performed an analysis using XAI method classes that have been successfully proven for AI applications. The goal was to find those methods that are most promising in terms of use for graders' support. These 5 most promising methods were then evaluated in a survey by graders. Due to the appropriate range of functions, the good display of our images with the XAI methods and its platform independence, we conducted the survey with Google Forms[2].

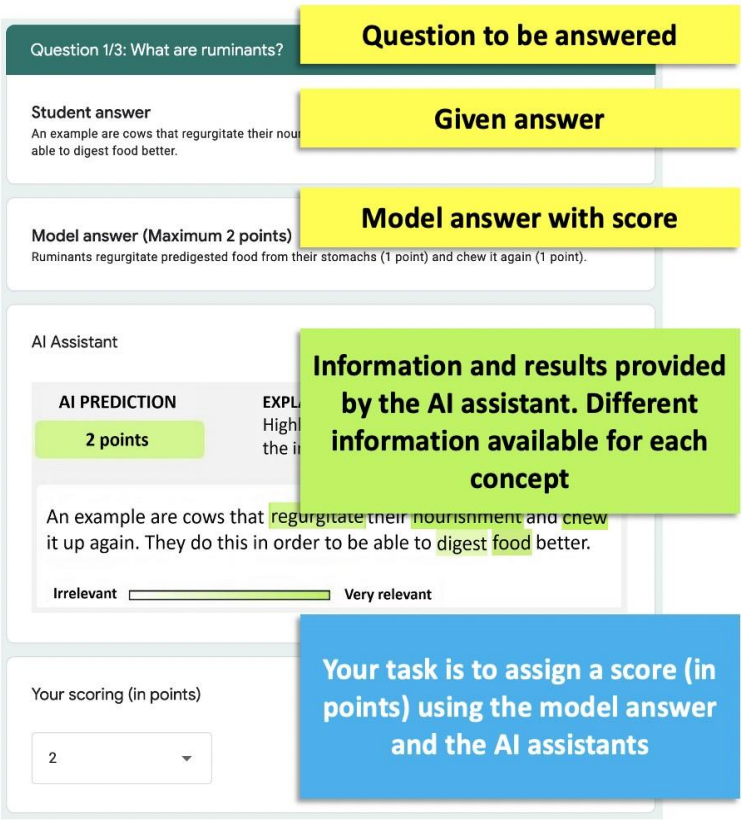

Fig. 8. Interface for grading in the questionnaire.

As demonstrated in Figure 8, the participants were asked to take the role of a teacher and evaluate various student answers given a model answer, a score predicted by the AI, and an explanation for the predicted score. In their role as graders, participants could assign 0, 1, or 2 points, with the highest possible score being 2 points. The type of explanation changed as the survey progressed. Before a new XAI method was displayed, questions were asked about the XAI method regarding the aspects *trust*,

[2] https://docs.google.com/forms

informative content, *speed*, *consistency and fairness*, *fun*, *comprehensibility*, *applicability*, *use in exam preparation*, and *in general*. The participants evaluated the questions regarding to the aspects with a score. The score range follows the rules of a forced choice Likert scale, which ranges from (1) *strongly disagree* to (5) *strongly agree*. Additionally, we evaluated the influence of the AI assistant's predictions on the graders' decisions by showing participants 2 correct and 1 incorrect point predictions of the AI assistant for each XAI method, i.e., one-third of the AI assistant's point prediction was incorrect.

71 participants (36 female, 35 male) filled out our questionnaire. The participants of our user study were professors, lecturers, and teachers between 24 and 65 years old who participated free of charge. Most are employed at our university. But some are professors and lecturers at other universities or teachers in schools. We appreciate these distributions as it was important to us to get feedback from different people.

5 Experiments and Results

In this section, we will describe the results of our study in which we examined the XAI methods with regard to the aspects *trust*, *informative content*, *speed*, *consistency and fairness*, *fun*, *comprehensibility*, *applicability*, *use in exam preparation*, and *in general*. In addition, we investigated how strong the influence of the XAI method is on the grader's score—when the AI's prediction is correct and when it is not.

5.1 Trust

We asked the participants in our questionnaire if they based their evaluation on the AI assistant. The goal was to find out how much *trust* the participants have on the AI assistant depending on the XAI method. Figure 9 illustrates the feedback on the trust. On average, the highest trust is on *Points+matching positions* (3.30), followed by *Confidence+similar answers* (2.94), *Confidence* (2.80), *Relevance of words* (2.71), and *Points* (2.46).

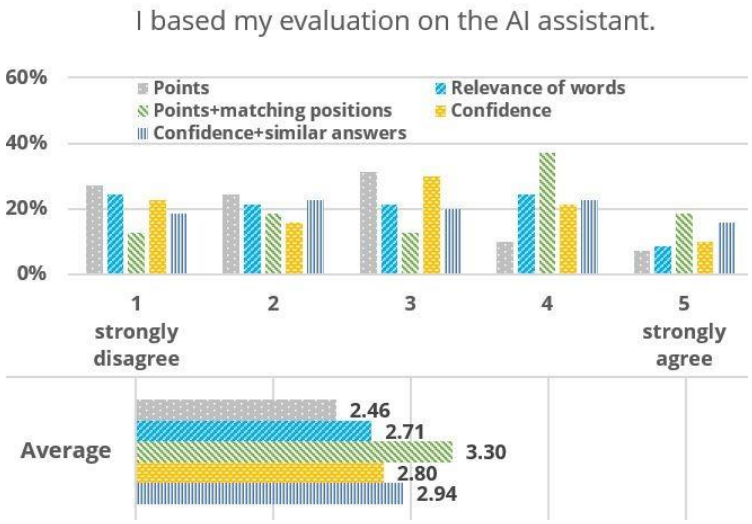

Fig. 9. Trust.

5.2 Informative Content

Figure 10 illustrates our evaluation in relation to the *informative content* of the suggested XAI methods. This time the averages are all above 3.00: *Points+matching positions* was rated on average with 3.57, *Confidence+similar answers* with 3.11, *Relevance of words* with 3.06, *Confidence* with 2.77, and *Points* with 2.60.

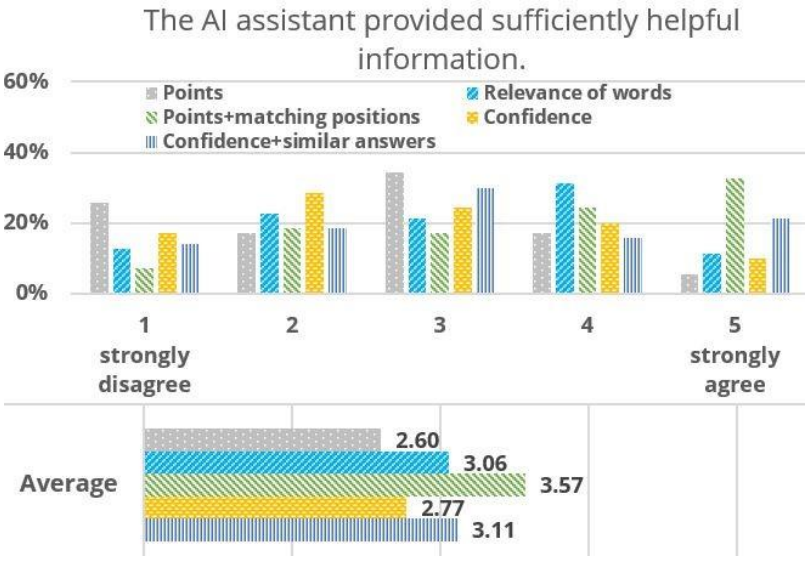

Fig. 10. Informative Content.

5.3 Speed

As illustrated in Figure 11, in the category *speed* with 3.70 *Points+matching positions* is again the best rated method. In comparison, the question if the AI assistant could help evaluate faster was rated only with an average score of 3.29 with *Relevance of words*, 3.17 with *Points*, 3.04 with *Confidence+similar answers*, and 3.03 with *Confidence*.

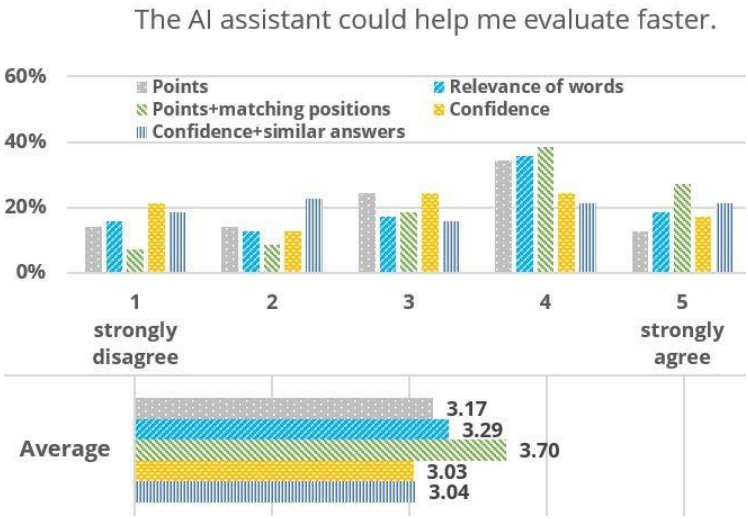

Fig. 11. Informative Content.

5.4 Consistency and Fairness

Then we asked the participants in our questionnaire if the AI assistant could help evaluate more *consistently and fairly*. The background to this is that graders do not always agree on the allocation of points and are also influenced in their grading by external factors that do not directly relate to the quality of the student answer [26]. The results are demonstrated in Figure 12. This time, the averages are closer together: The highest value remains at *Points+matching positions* (3.44), followed by *Confidence+similar answers* (3.31), *Relevance of words* (3.24), *Points* (3.21), and *Confidence* (3.10).

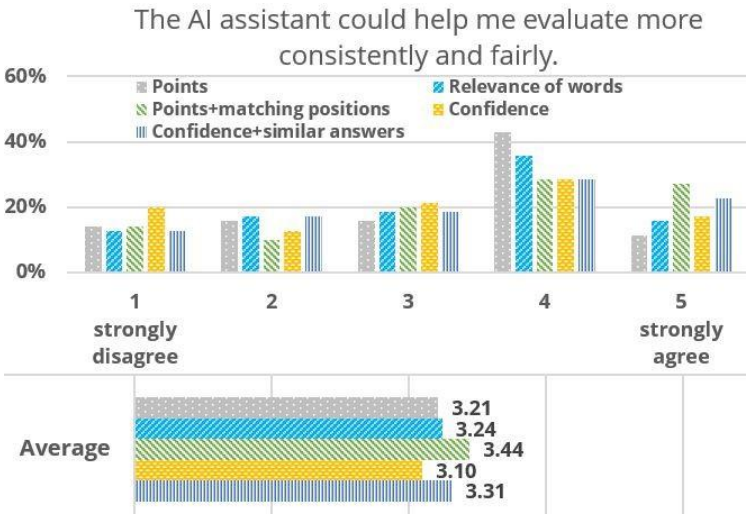

Fig. 12. Consistency and Fairness.

5.5 Fun

We asked the participants in our questionnaire if the rating with the AI assistant would be *fun* using the selected XAI methods. The results are shown in Figure 13. We see that on average, the highest value is again achieved by *Points+matching positions* (3.76), this time followed by *Relevance of words* and *Points* (3.59). Then comes *Confidence+similar answers* (3.26) and *Confidence* (3.16).

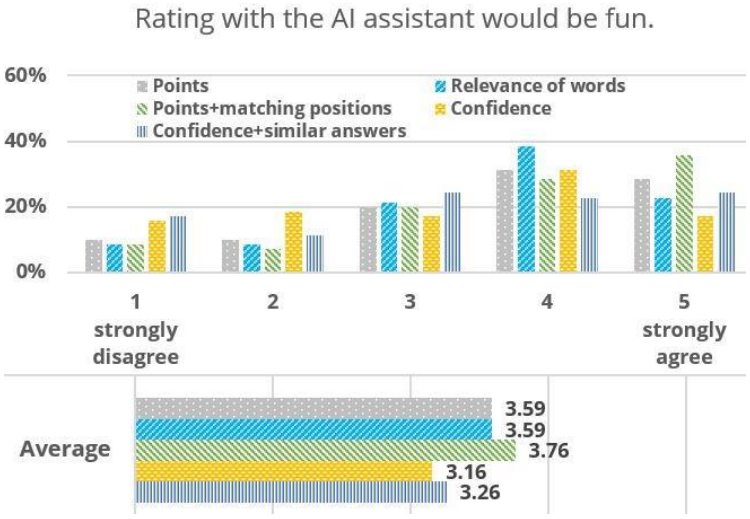

Fig. 13. Fun.

5.6 Comprehensibility

Another goal of our survey was to find out how comprehensible the XAI methods are. Therefore, our participants were asked if they were able to verify the recommendation of the AI assistant. Figure 14 illustrates the evaluation with regards to *comprehensibility*. Here, *Points+matching positions* (3.83), *Relevance of words* (3.76), and *Points* (3.61) are ahead on average. *Confidence+similar answers* (3.34) and *Confidence* (2.89) seem to be more difficult to understand on average.

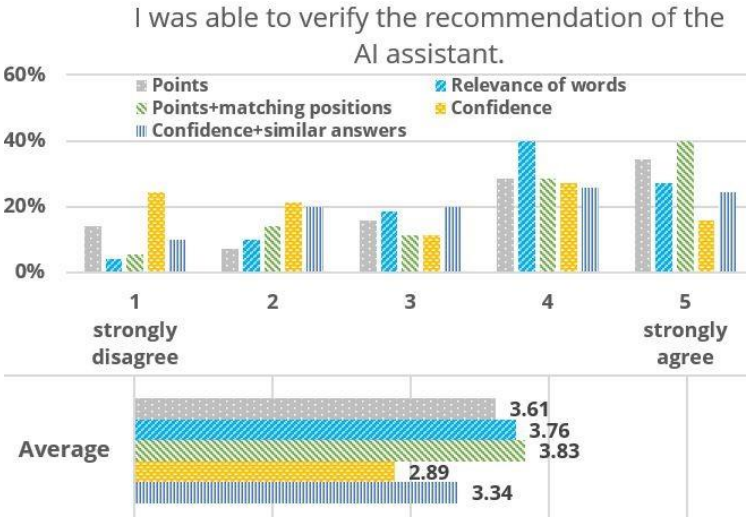

Fig. 14. Comprehensibility.

5.7 Applicability

The distribution in averages that we see in comprehensibility is also seen in *applicability* as shown in Figure 15: The highest value remains at *Points+matching positions* (3.79), closely followed by *Relevance of words* (3.67), and *Points* (3.59). *Confidence+similar answers* (3.31) and *Confidence* (3.11) make up the tail.

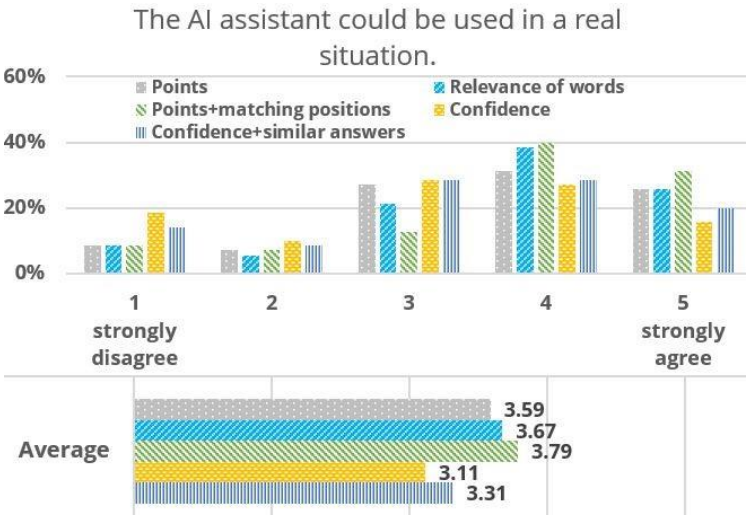

Fig. 15. Applicability.

5.8 Use in Exam Preparation

While the focus of the questions so far was on the use of XAI methods for the support of graders, XAI methods can also help learners prepare for an exam. Consequently, we asked our participants for each XAI method if they think that it is useful for learners as well. Again, *Points+matching positions* (3.41), *Relevance of words* (3.39) and *Points* (3.33) are close to each other. Again *Confidence+similar answers* (3.14) and *Confidence* (2.96) bring up the rear.

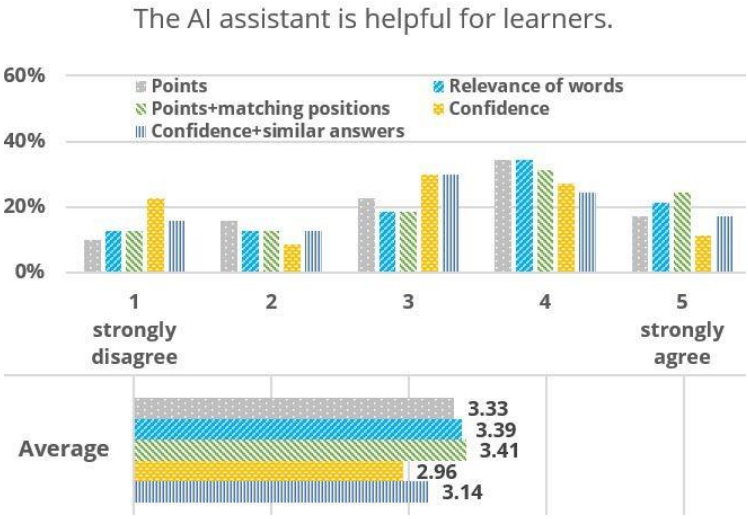

Fig. 16. Use in Exam Preparation.

5.9 General Evaluation

The last aspect we asked about was how the XAI methods are evaluated *in general*. As illustrated in Figure 17, the trends are comparable as in the other aspects: With 3.94 on average, *Points+matching positions* is rated as good. Then comes *Relevance of words* (3.57) with 10% less, followed by *Confidence+similar answers* (2.97), *Confidence* (2.80), and *Points* (2.54).

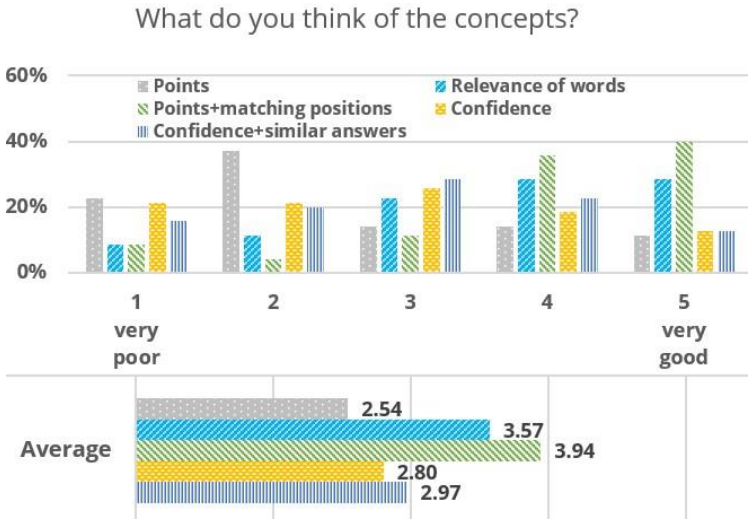

Fig. 17. Use in Exam Preparation.

5.10 Influence

Finally, we investigated how strong the *influence* of the XAI method is on the grader's score—when the AI's prediction is correct and when it is not. We evaluated the influence by showing participants 2 correct and 1 incorrect point predictions of the AI assistant for each XAI method, i.e., one-third of the AI assistant's point prediction was incorrect.

Figure 18 visualizes the percentage of correct scored student answers by the graders and the average deviations from the correct score in the case of a correct point prediction by the AI assistant and in the case of an incorrect point prediction by the AI assistant for all tested XAI methods. With a maximum score to be obtained per question of 2 points, the deviations of the graders range between 0.20 points (*Points+matching positions and relevance of words*) and 0.28 points (*Confidence*), which is only between 10% and 14%. For comparison, in the literature deviations of 15% between 2 graders are reported [17, 26], which shows that the tested XAI methods have no bad influence on the grading process. While in the case of a correct point prediction an average of 75% to 98% of the assessments achieved the same score as the assigned reference score, in the case of an incorrect point prediction an average of between 45% and 70% achieved the same score as the assigned reference score. For *points+matching*

positions, which performed best in the previous questions, 89% graded correctly in the case of correct point prediction and 63% in the case of incorrect point prediction. One must keep in mind here that in our study, one-third of the AI assistant's point prediction was incorrect.

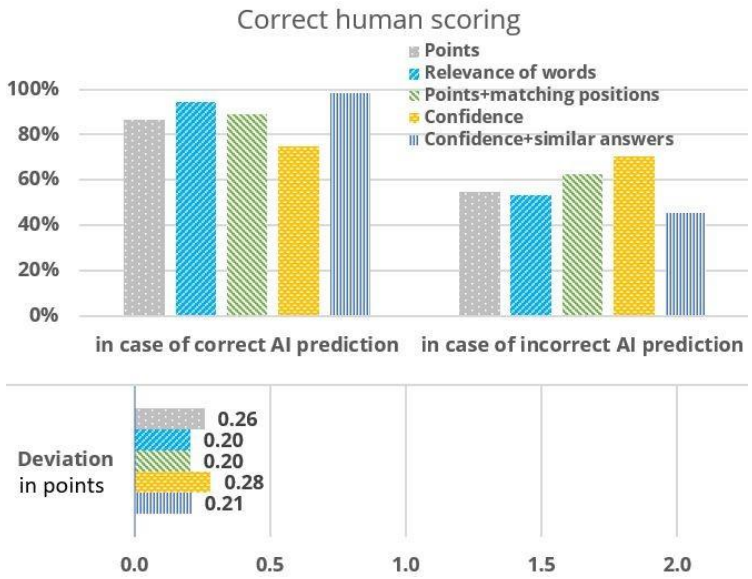

Fig. 18. Influence.

6 Conclusion

In this paper, we have investigated and evaluated different methods for explainability in ASAG. Our survey of over 70 professors, lecturers and teachers with grading experience showed that a clear majority of participants strongly agrees that it is important for them to understand how the AI reaches its expected scoring and their confidence in an AI grading support increases when it explains itself. Displaying the predicted points together with matches between student answer and model answer is rated better than the other tested XAI methods. Participants were asked if they agreed that the displayed XAI method helps for the aspects *trust*, *informative content*, *speed*, *consistency and fairness*, *fun*, *comprehensibility*, *applicability*, *use in exam preparation*, and *in general*.

Table 3 summarizes the average Likert scores for each evaluated aspect of the best method (*Points+matching positions*). Here the positive tendency for this method is shown as the scores are between 3 and 4, where 1 means *completely disagree* and 5 means *completely agree*. In addition, the relative improvement compared to the second-best method in each of the aspects is demonstrated. The statistical significance was tested with the type I error p = 2.5% with a Student's t-test for paired samples with n = 71.

Table 3. Average Likert Scores for *Points+matching positions* over the evaluated aspects and rel. improvement to 2nd-best XAI method.

Aspect	∅ Likert Score	Improvement over 2nd-best XAI method
trust	3.30	+12.2%* (*Confidence+similar answers*)
informative content	3.57	+14.8%* (*Confidence+similar answers*)
speed	3.70	+12.5%* (*Relevance of words*)
consistency & fairness	3.44	+ 3.9% (*Confidence+similar answers*)
fun	3.76	+ 4.7%* (*Relevance of words / Points*)
comprehensibility	3.83	+ 4.5% (*Relevance of words*)
applicability	3.79	+ 3.3%* (*Relevance of words*)
use in exam preparation	3.41	+ 0.6% (*Relevance of words*)
in general	3.94	+10.4%* (*Relevance of words*)

* statistically significant

Additionally, we investigated how strong the influence of the XAI method is on the grader's score—when the AI's prediction is correct and when it is not. The deviations of the graders from the actual points ranged between 10% and 14%. For comparison, in the literature deviations of 15% between 2 graders are reported [7,27], which shows that the tested XAI methods have only little influence on the overall grading process.

7 Future Work

Our goal was to survey a large representative group of teaching staff consisting of professors, lecturers and teachers with exam questions, student answers and model answers that are understandable for all participants. Therefore, we did not make an analysis of the individual lecturers' experience, subjects taught, performance at different difficulty levels, etc. This could be investigated in more detail in a future analysis.

Due to the very high performance of point prediction and the good results of the XAI methods in our survey, we plan to use ASAG together with the best XAI method at our university. In addition to grading, ASAG can also be used for exam preparation with an app or in online learning [28]. Consequently, future work may include to analyze the use of our XAI methods in interactive training programs to prepare students optimally for exams. Since in our study we considered the ASAG model as a black box model and produced explainability with another model, a graders' support by the direct interpretation of the complex ASAG models could be also investigated.

References

1. United Nations: Sustainable Development Goals: 17 Goals to Transform our World (2021), https://www.un.org/sustainabledevelopment/sustainable-development-goals

2. Correia, A.P., Liu, C., Xu, F.: Evaluating Videoconferencing Systems for the Quality of the Educational Experience. Distance Education 41, 4, 429–452. https://doi.org/10.1080/01587919.2020.1821607 (2020)
3. Koravuna, S., Surepally, U.K.: Educational Gamification and Artificial Intelligence for Promoting Digital Literacy. Association for Computing Machinery, New York, NY, USA (2020)
4. Chen, L., Chen, P., Lin, Z.: Artificial Intelligence in Education: A Review. IEEE Access 8, 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510 (2020)
5. Heffernan, N.: The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. International Journal of Artificial Intelligence in Education 24. https://doi.org/10.1007/s40593-014-0024-x (2014)
6. Libbrecht, P., Declerck, T., Schlippe, T., Mandl, T., Schiffner, D.: NLP for Student and Teacher: Concept for an AI based Information Literacy Tutoring System. In The 29th ACM International Conference on Information and Knowledge Management (CIKM2020). Galway, Ireland (2020)
7. Schlippe, T., Sawatzki, J.: Cross-Lingual Automatic Short Answer Grading. In Proceedings of The 2nd International Conference on Artificial Intelligence in Education Technology (AIET 2021). Wuhan, China (2021)
8. Adadi, A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052 (2018)
9. Ng, A.: Machine Learning Yearning. Online Draft. https://github.com/ajaymache/machine-learning-yearning (2017)
10. Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 (2017)
11. Hansen, L.K., Rieger, L.: Interpretability in Intelligent Systems – A New Concept? Springer, 41–49. https://doi.org/10.1007/978-3-030-28954-6_3 (2019)
12. Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., Rinzivillo, S.: Benchmarking and Survey of Explanation Methods for Black Box Models. arXiv:2102.13076 (2021)
13. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics 8, 8. https://doi.org/10.3390/electronics8080832 (2019)
14. Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., Sen, P.: A Survey of the State of Explainable AI for Natural Language Processing. arXiv:2010.00711 (2020)
15. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R.: Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. Proc. IEEE 109, 3, 247–278. https://doi.org/10.1109/JPROC.2021.3060483 (2021)
16. Rudin, C., Radin, J.: Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. Harvard Data Science issue 1.2. https://doi.org/10.1162/99608f92.5a8a3a3d (2019)
17. Sawatzki, J., Schlippe, T., Benner-Wickner, M.: Deep Learning Techniques for Automatic Short Answer Grading: Predicting Scores for English and German Answers. In Proceedings

of The 2nd International Conference on Artificial Intelligence in Education Technology (AIET 2021). Wuhan, China (2021)
18. Burrows, S., Gurevych, I., Stein, B.: The Eras and Trends of Automatic Short Answer Grading. International Journal of Artificial Intelligence in Education 25, 60–117 (2014)
19. Camus, L., Filighera, A.: Investigating Transformers for Automatic Short Answer Grading. Artificial Intelligence in Education 12164, 43–48 (2020)
20. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR. arXiv:1907.11692 (2019)
21. Pires, T., Schlinger, E., Garrette, D.: How Multilingual is Multilingual BERT?. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, 4996–5001. https://doi.org/10.18653/v1/P19-1493 (2019)
22. Poulton, A., Eliens, S.: Explaining Transformer-Based Models for Automatic Short Answer Grading. In Proceesings of the 5th International Conference on Digital Technology in Education (ICDTE 2021). Association for Computing Machinery, New York, NY, USA, 110–116. https://doi.org/10.1145/3488466.3488479 (2021)
23. van der Waa, J., Schoonderwoerd, T., van Diggelen, J., Neerincx, M.: Interpretable Confidence Measures for Decision Support Systems. International Journal of Human-Computer Studies 144, https://doi.org/10.1016/j.ijhcs.2020.102493 (2020)
24. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Francisco, California, USA) (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778 (2016)
25. Kim, B., Wattenberg, M., Gilmer, J., Cai, C.J., Wexler, J., Viégas, F., Sayres, R.: Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In ICML 2018
26. Hanna, R.N., Linden, L.L.: Discrimination in Grading. American Economic Journal: Economic Policy 4, 4 (2012), 146–168. http://www.jstor.org/stable/23358248 (2012)
27. Mohler, M., Bunescu, R., Mihalcea, R.: Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Portland, Oregon, USA, 752–762 (2011)
28. Schlippe, T., Sawatzki, J.: AI-based Multilingual Interactive Exam Preparation. In The Learning Ideas Conference 2021 (14th annual conference). ALICE - Special Conference Track on Adaptive Learning via Interactive, Collaborative and Emotional Approaches. New York, USA. https://doi.org/10.1007/978-3-030-90677-1_38 (2021)