ExASAG: Explainable Framework for Automatic Short Answer Grading

Maximilian Törnqvist
Dept. of Computer and Systems Sciences,
Stockholm University
maximilian.tornqvist@dsv.su.se

Mosleh Mahamud
Dept. of Computer and Systems Sciences,
Stockholm University
mosleh.mahamud@dsv.su.se

Erick Mendez Guzman
Dept. of Computer Science,
Manchester University
erick.mendezguzman@manchester.ac.uk

Alexandra Farazouli
Dept. of Education,
Stockholm University
alexandra.farazouli@edu.su.se

Abstract

As in other NLP tasks, Automatic Short Answer Grading (ASAG) systems have evolved from using rule-based and interpretable machine learning models to utilizing deep learning architectures to boost accuracy. Since proper feedback is critical to student assessment, explainability will be crucial for deploying ASAG in real-world applications. This paper proposes a framework to generate explainable outcomes for assessing question-answer pairs of a Data Mining course in a binary manner. Our framework utilizes a fine-tuned Transformer-based classifier and an explainability module using SHAP or Integrated Gradients to generate language explanations for each prediction. We assess the outcome of our framework by calculating accuracy-based metrics for classification performance. Furthermore, we evaluate the quality of the explanations by measuring their agreement with human-annotated justifications using Intersection-Over-Union at a token level to derive a plausibility score. Despite the relatively limited sample, results show that our framework derives explanations that are, to some degree, aligned with domain-expert judgment. Furthermore, both explainability methods perform similarly in their agreement with human-annotated explanations. A natural progression of our work is to analyze the use of our explainable ASAG framework on a larger sample to determine the feasibility of implementing a pilot study in a real-world setting.

1 Introduction

Assessment is fundamental to any educational process as an evaluation system reflecting individual performance and a way to compare results across populations (Harlen et al., 1992). Two key elements to consider when designing an assessment are question type and grading method (Gardner, 2012). While questions may come in various forms, such as multiple-choice questions, short answers, or essays, the grading method can be either manual grading performed by domain experts or automatic grading by computational methods (Broadfoot and Black, 2004).

Previous research has established that assessing free-text short answers is a process that, besides being time-consuming, may lead to inequalities due to the difficulties in applying consistent evaluation criteria across answers (Page, 1994; Gardner, 2012). Data from several studies suggest that teachers dedicate approximately 25% to 30% of their time grading written examinations (Broadfoot and Black, 2004; Sukkarieh et al., 2003). Moreover, manual grading requires concentration for long periods of time, which could lead to differences in grading for answers with similar quality, creating inequities in the assessment process and its outcome (Whittington and Hunt, 1999; Burrows et al., 2015).

In the literature, automatic short answer grading (ASAG) is defined as the task of assessing short natural language responses to objective questions using computational methods (Page, 1994; Whittington and Hunt, 1999). ASAG techniques have evolved from traditional rule-based models to state-of-the-art systems utilizing deep learning-based natural language processing (NLP) models (Sukkarieh et al., 2003; Leacock and Chodorow, 2003; Galhardi and Brancher, 2018). Researchers have been able to build supervised learning models based on assessment questions, answers provided by students, and the corresponding grades assigned by teachers (Burrows et al., 2015; Willis, 2015). The objective is, therefore, to predict which label score a new question-answer pair should achieve.

Over the past five years, researchers have leveraged the power of novel deep learning architectures such as the Transformer (Vaswani et al., 2017) to

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 361–371
July 13, 2023 ©2023 Association for Computational Linguistics

improve accuracy for ASAG models (Sung et al., 2019a). Nevertheless, the performance improvement has come at the cost of models becoming less understandable for stakeholders, and their opaqueness has become an obstacle to their deployment in the educational domain (Belle and Papantonis, 2020; Arrieta et al., 2020). Consequently, Explainable Artificial Intelligence (XAI) has emerged as a relevant research field aiming to develop methods that allow stakeholders to understand the outcome of deep learning-based systems (Gunning et al., 2019; Arrieta et al., 2020). As such, several lines of evidence suggest that providing insights into models' inner workings might be helpful in building trust in these systems and detecting potential biases (Belle and Papantonis, 2020; Arrieta et al., 2020; Jacovi and Goldberg, 2021).

A great deal of previous research into XAI methods for explaining NLP models has focused on building reliable associations between the input text and output label and quantifying how much each element (e.g., word or token) contributes to the final prediction (Danilevsky et al., 2020). Such XAI methods can usually be divided into feature importance-based explanations (Simonyan et al., 2013), perturbation-based explanations (Zeiler and Fergus, 2014), explanations by simplification (Ribeiro et al., 2016) and language explanations (Lei et al., 2016). Previous studies have indicated that *rationales* or language explanations are easier to understand and use since they are verbalized in human-comprehensible natural language (Lei et al., 2016; DeYoung et al., 2019).

This study focuses on explaining binary text classification for student responses gathered from a Data Mining course exam. As such, the main objective is to generate a framework that predicts binary grades and simultaneously produces associated rationales in order to justify the predicted grade of a given student response. By doing so, we intend to enrich the insights given by previous research, by presenting a framework that demonstrates how recent progressions of deep learning architectures and XAI can be combined in order to address the problem of ASAG. As such, we aim to set an example for how future research can incorporate XAI in the educational domain. Conclusively, our main contributions are as follows:

1. Suggesting a framework for creating sentence-level and word-level attributions by utilizing token-level relevancy scores.
2. Evaluating contemporary explainability methods by measuring the Intersection-Over-Union of our language explanations and human rationales.
3. Applying a fine-tuned Transformer model to perform ASAG on data-scientific question-answer pairs by utilizing collected data from a course in Data Mining.

2 Related Work

Large Language Models (LLMs) such as Transformer models have been increasingly applied in the domain of ASAG (Haller et al., 2022). Given a limited amount of examples, Transformer models such as BERT have proven their capability to achieve state-of-the-art performance within the field of ASAG (Sung et al., 2019b). The ability to handle single short documents, such as question-answer pairs, makes BERT a suitable model for various downstream tasks (Devlin et al., 2018). Most previous research and implementations focus on the model's effectiveness using standard classification metrics such as F1 and accuracy, precision, and recall (Haller et al., 2022). However, there is a limited amount of research addressing *why* certain predictions are being made. As a consequence, a lack of trust and understanding of the model predictions remains an issue. Thus, our work explores the use of explainability techniques as a tool for ASAG, in order to increase the understanding of the predictions being done.

Rationale extraction refers to a post-hoc explainability method for NLP models in which the goal is to create deep learning-based NLP solutions explainable by uncovering part of an input sequence that the prediction relies on the most (Lei et al., 2016; DeYoung et al., 2019). Most previous research on rationale extraction has been carried out using an *encoder-decoder* architecture. In such a setting, the *encoder* works as a tagging model, where each word in the input sequence receives a binary tag indicating whether it is included in the rationale. The *decoder* then only accepts the input highlighted as a rationale and maps it to the target labels (Zaidan et al., 2007; Bao et al., 2018; Narang et al., 2020).

Previous studies have proposed a multi-task learning approach for rationale extraction utilizing two models and training them jointly to minimize a composite cost function (Lei et al., 2016; Bastings et al., 2019; Paranjape et al., 2020). Unfortunately, one of the main drawbacks of multi-task learning architectures for rationale extraction is that it is challenging to train the encoder and decoder jointly under instance-level supervision (Zhang et al., 2016; Jiang et al., 2018). Pipelined models are a simplified version of the encoder-decoder architecture in which the encoder is first trained to extract the rationales. Then the decoder is fit to perform prediction using only the rationale (Zhang et al., 2016; Jain et al., 2020). It is important to note that no parameters are shared between the two models and that rationales extracted based on this approach have been learned in an unsupervised manner since the encoder is deterministic by nature.

There is little consensus on what makes a good machine-generated rationale and how to evaluate a rationale for benchmarking. Most researchers investigating rationale evaluation have utilized *proxy-based* methods, where rationales are assessed based on automatic metrics that attempt to measure desirable properties (Carton et al., 2020). One of the most common methods for evaluating rationales is to measure how well they agree with explanations provided by human annotators (DeYoung et al., 2019). In the context of explainable NLP, this property is referred to as *plausibility*. As such, it is usually evaluated based on the token overlap between human annotations and machine-generated rationales. Using such an approach, researchers have been able to derive token-level precision, recall, and F1 scores using Intersection-over-Union (IOU) at token level (Paranjape et al., 2020; Chan et al., 2021; Guerreiro and Martins, 2021).

3 Explainable Autograding Framework

The explainable framework is illustrated in Figure 1, consisting of an encoder responsible for generating explanations and a decoder responsible of performing the binary classification.

3.1 Encoder

The encoder is built using two main components, where the first component corresponds to the explainability method of use, and the second component corresponds to the ranking and processing of the given attributions created by the used explainability method. The two mentioned components result in a ranking for each sentence in a student's answer based on its importance in end classification. Thus, the concept of the framework itself is not dependent on the individual explainability methods presented in this study. As such, with minor adjustments according to the outputs of the used method, the concept of the presented framework should be considered generalizable and possible to implement in conjunction with other token-based methods of attribution. Subsequently, the following paragraphs will introduce the explainability methods being used in this study.

3.2 Explainability methods

As the complexity of a model increases, the model itself cannot longer be used as a method for explanation. As such SHAP utilizes cooperative game theory and Shapley values to explain a model's output prediction (Lundberg and Lee, 2017). By doing so, SHAP creates an interpretable approximation of the original model, which is refered to as the explanation model. In essence, SHAP is a model-agnostic explainability method that captures the importance value of an input feature by perturbating the input feature and observing the change in the model's prediction output. By observing the resulting output of the perturbation, SHAP makes it possible to assign each input feature an importance value. In practice, SHAP utilizes additive feature attributions, which in essence can be defined as a mapping of the original input features to simplified features. As such, it achieves the aforementioned interpretable approximation of the original model. In the task of ASAG, the tokens included in the answer correspond to the input features. Consequently, each and every token in the answer will be given a relevancy value.

Similarly, each input feature is assigned an attribution value with Integrated Gradients (IG). IG is an explainability method based on two main axioms; Sensitivity and Implementation Invariance (Sundararajan et al., 2017). IG measures the attribution value by comparing the model's output function of the input with the model's output function of an uninformative baseline. The uninformative baseline could correspond to a black image in an object recognition task, while for text classification, the baseline could correspond to a zero embedding vector. The integrated gradients can then be defined as 'the path integral of the gradients along the straight line path from the baseline to the input' (Sundararajan et al., 2017). For a text classification task, the integrated gradients are calculated by interpolating between the baseline and the original output for *k* number of steps. This

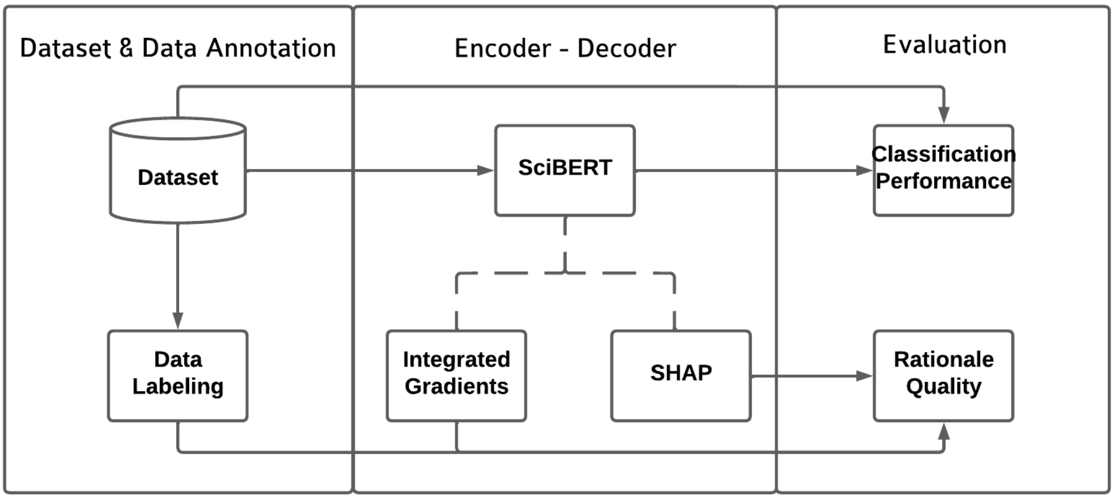

Figure 1: Explainable Framework for Automatic Short Answer Grading.

gives insight into each feature independent of the others and its impact on the output prediction. Furthermore, Gradient-based explanations are known to be robust and efficient (Nielsen et al., 2022).

The reason why SHAP and IG are used in this experiment is that SHAP can handle interactions between features when explaining (Nielsen et al., 2022). In contrast, IG only considers each feature's individual contribution, making it suitable to observe both effects. Both post-hoc explainability methods have a reputation for being robust, and neither has an effect on the end classification accuracy (Vale et al., 2022; Lakkaraju et al., 2020) hence, making it reasonable to apply to ASAG tasks.

3.2.1 Sentence Level Explanations

Since Transformer models usually represent singular words as multiple tokens, explainability methods such as SHAP and IG will return attributions at a token level when used in combination with Transformers. In this framework, attributions are grouped per sentence, creating Sentence Level Attributions (SLAs). The SLAs are all based on Word Level Attributions (WLAs), which in turn are based on the original Token Level Attributions (TLAs) generated by the explainability methods.

We define the WLAs as the sum of all the TLAs representing a single word. Furthermore, as including stopwords in the SLA could lead to very neutral attribution values of sentences with a considerable amount of stopwords, we define the SLAs as the mean attribution of all non-stopwords contained in a sentence. As such, stopwords are assumed not to be highly determining for the end classification. Thus, they are completely ignored in the calculation of SLA. Furthermore, as a consequence of the partially arithmetic characteristics of the data set and Transformers' inability to handle such arithmetics, any non-alphabetical characters are removed before calculating the SLAs.

3.3 Decoder

The components of the decoder are a Transformer model fine-tuned on a exams from a data mining course, where the characteristics of the data set are further detailed in section 4.1.

The model For the classification of the text, we use SciBERT (Beltagy et al., 2019), as it is a pre-trained model based on the architecture of BERT, which uses a corpus of 1.14 million papers instead of the original pre-training data found in BERT. Of these 1.14 million papers, 18% of the papers in the corpus comes from the domain of computer science. In terms of representing language, the vocabulary of the SciBERT model only overlaps the vocabulary of the BERT model by 42% (Beltagy et al., 2019). As such, this difference in vocabulary illustrates the differences between scientific text in comparison to general text. Furthermore, it also highlights the importance of choosing the appropriate model and associated vocabulary depending on the domain of the given task. Given that SciBERT can be considered to be of a somewhat computer scientific

3.4 Evaluation

Classification Performance For evaluating the classification performance of the used model, we calculate precision, recall, and F-1 score against the given test set.

Plausibility For calculating the quality of the model rationales, we calculate the Inter-Annotator Agreement (IAA) (Zaidan et al., 2007; Carton et al., 2020) as the Intersection-Over-Union (IOU). The use of IOU is to calculate the overlap of the model rationales and the human rationales. Before calculating this overlap, we first ensure that the model rationales and the human rationales are in a comparable format. To achieve this, both of them are processed in a standardized manner.

As the explanation models used generates attribution scores for all text in the given response, the generated attributions needs to be filtered in order to conduct a fair comparisson with human rationales (as they do not usually contain all of the text in a given response). As such, the sentences can be ranked by their respective attribution value. Following such a ranking, the top *k* attributions corresponding to the given grade could be picked out for comparisson with the human rationales, where *k* is defined by the number of rationales annotated by the human. As such, the top *k* sentences with the highest SLA are selected for comparison with the human rationales if the label is “Satisfactory”, where *k = the number of sentences in human rationales*. However, if the label of the answer is "Non-satisfactory", the *k* sentences with the lowest SLA are selected for comparison with the human rationales.

In order to compare the sentences, both set of sentences are split into non-stopword tokens. From these sets of tokens, empty strings and non-alphabetic characters are removed. Finally, the two sets of tokens will represent the model and human rationales when calculating the IAA.

4 Data, Experiments and Results

4.1 Data set

As part of a project in automatically grading exams at Stockholm University, the data selected in this study was selected in order to partly evaluate the potential of using automatically grading systems on low-resource data. As such, the selected data set used in this experiment is an English data set consisting of 1131 question-answer pairs collected from graded exams of a Data Mining course at Stockholm University. As such, the data has been collected from a limited amount of course iterations. In total, there are 31 unique questions, with an average of 36,5 answers per question. Given the amount of question answers pairs, the adjustments and changes that have been applied to the questions inbetween the given iterations, and the amount of answers per question, it is reasonable to deem the data set to be of a low-resource charachter. In essence, this poses a fundamental challenge for building grading systems, where the amount of examinatory data can be limited due to a multitude of factors such as limited data collection, frequent adjustments to questions or course content, or due to the course being new. As such, utilizing such a data set, will help evaluate the potential of building automatically grading systems on low-resource data.

The data set also features a lot of scenario-based questions, where the student is often asked to provide a solution for a scenario-based problem. This type of response generally involves complex reasoning about the problem and as a consequence, the answers are usually long compared to answers in data sets previously used, with an average length of 155 words per answer across the whole data set. Given this, it could be argued that the task of grading these answers could be seen as a more elaborate version of the ASAG task that a lot of previous research has been focusing on (Haller et al., 2022). Furthermore, some of the question-answer pairs involve small amounts of arithmetics. Given the amount of available data and varying class representation, the scales of grading have been converted from the original scales (0-5, 0-8, and 0-10) to binary labels (0-1). From the original scales, binary labels were derived by assessing every answer that achieved 50% or more of the original maximum grade as a satisfactory(1) answer and every answer that achieved less than 50% of the original maximum grade as a non-satisfactory answer(0). Following the conversion, there are 667 satisfactory answers and 464 non-satisfactory answers.

4.2 Data annotation

Before performing the annotation, we developed an annotation scheme and guidelines to facilitate labeling question-answer pairs (Krippendorff, 2004). The scheme is based on the rubric associated with each question defined by Stockholm University lecturers. As mentioned before, we focused on binary text classification. Consequently, we asked our annotators to label each item as “Satisfactory” or “Non-Satisfactory” based on whether they would assign at least 50% or more of the total maximum grade for each question. To illustrate, an answer graded 10 points to a question worth 20 points would be satisfactory, while an answer graded 9 points to the same question would be labeled as non-satisfactory. However, since our goal is to provide richer annotations that support grading, we also asked our annotators to select phrases and sentences to justify their labeling decisions (Zaidan et al., 2007; DeYoung et al., 2019; Guzman et al., 2022). The annotation guidelines and examples of our dataset are available upon request.

Since annotations of the original dataset was not available, the annotation of the corpus was completed by two annotators aged above 25 years old with degrees in Data Mining and Computer Science from Stockholm University. Considering how domain-specific our research is and the data privacy constraints of our dataset, we decided against crowd-sourcing the annotation. During the annotation process, the annotators were encouraged to ask questions over online sessions to facilitate feedback and ensure high-quality human rationales (Nowak and Rüger, 2010). In order to avoid any bias or preconceptions being passed on from the authors to the annotators during the feedback sessions, the annotations were carried out prior to the creation of any model rationales. Furthermore, in order to avoid being directly involved in any of the examples, we highly encouraged the annotators to ask questions of a conceptual character rather than to showcase specific examples from the dataset.

To validate our annotation guidelines, we randomly selected 20 question-answer pairs and asked our annotators to label them independently using LightTag (LightTag, 2018) as the annotation platform. This preliminary validation helped the annotators to familiarize themselves with the scope of the task and to understand how to use LightTag. The trial run enabled us to obtain constructive feedback on the annotation scheme and guidelines (Zou et al., 2021).

We assessed the quality of the annotations using the F1 score as IAA metric (Zaidan et al., 2007; Carton et al., 2020). Considering the aim of our research, we computed IAA at the level of binary labels and rationales (Krippendorff, 2004). Considering the annotations of our most senior annotator (A1) as the gold standard, we obtained a micro-averaged F1 score of 0.94 for the 20 items in the trial run.

As mentioned before, measuring exact matches between rationales is likely too strict. Similarly to what we described as one of the evaluation metrics for the encoder, we used IOU at a token level (DeYoung et al., 2019). For rationales’ IAA, the IOU is the size of the token overlap of the two human-generated explanations, divided by the size of their union (Carton et al., 2020). We counted it as a match if the IOU exceeds a user-defined threshold. Following (Zaidan et al., 2007), we utilized 0.5 as the threshold and derived a micro-averaged F1 score of 0.81 for rationales in the trial run.

Several lines of evidence suggest that reaching a high IAA for rationale labeling is still challenging, mainly because of the complexity of the annotation task itself and the subjective nature of the human rationales (Lei et al., 2016; Strout et al., 2019; Carton et al., 2020). Nevertheless, we observed a fair agreement between our annotators compared with previous work on rationales for binary text classification (Zaidan et al., 2007; DeYoung et al., 2019). Consequently, we sampled 200 items from our dataset and asked each annotator to label 100 question-answer pairs to consolidate the rationale-annotated dataset to evaluate our explainable framework.

Our annotators labeled almost two-thirds of the 200 question-answer pairs as “Satisfactory” (134 items). The human rationales for the “Satisfactory” label were, on average, 55 words-length with a standard deviation of 12 words. The rationales assigned to the “Non-Satisfactory” class were slightly shorter, with an average of 48 words and a standard deviation of 18 words.

4.3 Experiments

For the classification experiment, the data was split using stratification into a training set consisting of 757 examples and a test set consisting of 374 examples. Using the training and test set, the model was evaluated both with fine-tuning on the training set and without any fine-tuning. The aim of this method is to demonstrate the difference that fine-tuning can make in classification performance when the amount of data is limited (for results with no fine-tuning, see Appendix A).

Given the previously described question-answer pairs, the models were fine-tuned for 3 epochs with a batch size of 8. For optimization, AdamW was used with a learning rate of 2e-5 and a weight decay of 0.01.

For evaluating the performance of the classification model, a total amount of 1131 question-answer pairs were used. From these 1131 examples, 757 examples were used for fine-tuning the model, while 374 examples were used for testing the model. The metrics for measuring the performance of the classification model were precision, recall, and F1-score on a micro-level as well as on a macro-level, for both of the labels, which can be seen in Table 1.

For evaluating the performance of the explainability framework, a sample of 5 questions was chosen for this experiment. The sampling of questions was based on factors such as label distribution, the average length of answers, the number of answers per question, and the amount of arithmetics involved in the question. Since the data set was very limited in terms of the number of answers per question, we made sure that both of the class labels were represented in each of the sampled questions. Having this in mind, we also made sure not to include questions that were relatively high in arithmetical answers. The support of the individual questions ranges from 35 answers per question to 50 answers per question, with a mean of 41 answers per question. In total, the selected data set for evaluating the sentence explainability framework consisted of 200 question-answer pairs. Thus, given the limited annotation budget of the project, the explainability framework is only evaluated on a subset of the data set used for evaluating the classification task. As such, the questions with the most lengthy answers were also rejected as a part of the evaluation process.

5 Results

5.1 Classification results

Table 1 shows the classification performance of the model used in the explainability experiments, where the classification performance is evaluated using precision, recall, and F1-score. As seen in the table, there is a difference in classification performance between the two given labels. The difference in performance could be expected as a consequence of the imbalance in the data set.

	Precision	Recall	F1-score
Label 0	0.74	0.67	0.70
Label 1	0.79	0.84	0.82
Macro Avg	0.77	0.76	0.76

Table 1: Overall classification performance metrics of fine-tuned SciBERT, where Label 0 = Non-satisfactory and Label 1 = Satisfactory.

Table 2 shows F1-score and recall based on a varying threshold and the number of matches between the human rationales and the model rationales generated by IG. Where a match is registered if the IAA calculated as the IOU between the model rationales and the human rationales exceeds the given threshold. As mentioned in section 3, the calculation is carried out using two sets of tokens representing the human and model rationales. In this scenario, the ground truth will always be a match, which means that the recall will represent the number of matches made out of all possible matches. Given a Threshold of 0.5, the results show an F1-score of 0.62 and a recall of 0.45. This means that out of all possible matches, the IAA exceeds the 0.5 threshold in 45% of all answers.

Threshold	F1	Recall
0.1	0.95	0.91
0.2	0.92	0.85
0.3	0.82	0.70
0.4	0.75	0.60
0.5	0.62	0.45

Table 2: Overall performance metrics for IG, based on a threshold and the number of matches.

Table 3 shows the F-1 score and Recall based on a varying threshold and the number of matches between the human rationales and the model rationales generated by SHAP. If the IAA calculated as the IOU exceeds the threshold of 0.5 for a given answer, we calculate it as a match. Given a Threshold of 0.5, the results show an F1-score of 0.63 and a recall of 0.46. Which is similar to the results achieved by IG. This means that out of all possible matches, the IAA exceeds the 0.5 threshold in 46% of all answers.

6 Discussion

When comparing the F1-score and recall of the SHAP method with the F1-score and recall of the IG method there seems to be little to no difference in their respective IAA with the human annotators. However, both of the methods seem to do well given the complexity of the data as well as the lim-

Threshold	F1	Recall
0.1	0.96	0.92
0.2	0.89	0.81
0.3	0.83	0.70
0.4	0.77	0.62
0.5	0.63	0.46

Table 3: Overall performance metrics for SHAP, based on a threshold and the number of matches.

ited amount of data that was used for fine-tuning. Given the SciBERT model and these accompanying explainability methods, it seems to be possible to generate representative explanations as well as explanations that could be valuable for a human annotator.

Given that the data set used is not only considerably smaller but also considerably more complex in terms of answer length than most data sets previously used in the task of ASAG, a slight decrease in classification performance is expected compared to previous research. Furthermore, one implication of the classification results is that Transformer models seem to require a very small amount of question-specific data in order to substantially improve its performance in classification, even when given relatively complex data. However, such solutions may not replace human expertise. Rather, using a combination of these models and the presented explainability methods, this performance can increase confidence in the given explanations and as a consequence, it could help aid and assist human experts in grading when data is very limited.

7 Conclusion and Future Work

NLP tools hold immense potential for scoring free-text answers from students and augmenting teachers' evaluation capabilities in a scalable manner. Transformer-based models can help identify patterns from students' responses and prioritize solutions that need further checking. However, their black-box nature becomes an obstacle when deploying these models in real-world educational applications. To bridge this knowledge gap, we introduce an explainable ASAG framework that produces competitive predictions along with human-understandable natural language explanations. Our framework leverages LLMs capabilities combined with post-hoc explainability methods that do not require training, reducing the number of question-answer pairs needed to achieve state-of-the-art results.

Furthermore, the classification performance proves that LLMs can achieve competitive ASAG performance on complex questions with a low number of answers per question when given domain-specific training, indicating a low threshold for applying domain-specific ASAG. As a consequence, the resulting performance could give a certain degree of confidence when assisting teachers with valuable explanations.

Further work needs to be done to establish whether incorporating human-generated rationales during training can boost the model's predictive performance and the quality of its generated explanations (Strout et al., 2019; Lei et al., 2016). Our future work aims to incorporate them using a multi-task learning approach and evaluate rationales beyond the plausibility dimension covered in the presented article.

Finally, we hope our framework and initial results can help promote research on explainability in ASAG systems.

8 Limitations

Given that the WLAs are calculated as the sum of all the TLAs representing one single word, it is possible that there could be an underlying preference for longer words in the framework. However, multiple tokens in a word could also have conflicting attributions, so it is not entirely clear how this affects the framework. Given the results of this implementation, it could be reasonable to try and calculate the WLAs as the mean of all TLAs instead.

Furthermore, it is reasonable to discuss the consequences of the preprocessing steps being carried out in the experiment. Although such preprocessing steps might increase the IAA measured between the human rationales and model rationales, it is reasonable to question what these preprocessing steps actually result in and their possible value in real-world applications. In cases where the use case is to identify and highlight certain important words, such preprocessing steps might bring a considerable amount of value. However, if the end goal is to represent the model's attention as precisely as possible, these preprocessing steps might skew the representation of the model's attention. Consequently, one could argue that there exists a trade-off between usable model explanations, which can be used as an assisting or guiding tool for the human expert, and explanations that are fair representations of the model's inner workings. In the case of ASAG, explanations such as the ones created by the presented framework could likely be used as an assisting tool in helping human expert graders find important words or sentences. Given such a framework, the speed of grading could likely be increased without removing the trust of having a human grader making the end decision.

Lastly, it is worth noting that the use of top k sentences should only be seen as a means of calculating IAA. However, in a real-world inference setting, the number of relevant sentences might be dependent on the task as well as the subject. In the case of assisting a human expert in grading, the number of top k sentences might be a parameter controlled by the human expert in order to showcase only the most relevant sentences marked by the model annotations, where the number of relevant sentences might be dependent on the length of the student answer as well as the complexity of the given question.

References

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. *arXiv preprint arXiv:1808.09367*.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. *arXiv preprint arXiv:1905.08160*.

Vaishak Belle and Ioannis Papantonis. 2020. Principles and practice of explainable machine learning. *arXiv preprint arXiv:2009.11698*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Patricia Broadfoot and Paul Black. 2004. Redefining assessment? the first ten years of assessment in education. *Assessment in Education: Principles, Policy & Practice*, 11(1):7–26.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. *arXiv preprint arXiv:2010.04736*.

Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, and Hamed Firooz. 2021. Unirex: A unified learning framework for language model rationale extraction. *arXiv preprint arXiv:2112.08802*.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Lucas Busatta Galhardi and Jacques Duílio Brancher. 2018. Machine learning approach for automatic short answer grading: A systematic review. In *Ibero-american conference on artificial intelligence*, pages 380–391. Springer.

John Gardner. 2012. *Assessment and learning*. Sage.

Nuno Miguel Guerreiro and André FT Martins. 2021. Spectra: Sparse structured text rationalization. *arXiv preprint arXiv:2109.04552*.

David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. Xai—explainable artificial intelligence. *Science Robotics*, 4(37).

Erick Mendez Guzman, Viktor Schlegel, and Riza Batista-Navarro. 2022. Rafola: A rationale-annotated corpus for detecting indicators of forced labour. *arXiv preprint arXiv:2205.02684*.

Stefan Haller, Adina Aldea, Christin Seifert, and Nicola Strisciuglio. 2022. Survey on automated short answer grading with deep learning: from word embeddings to transformers. *arXiv preprint arXiv:2204.03503*.

Wynne Harlen, Caroline Gipps, Patricia Broadfoot, and Desmond Nuttall. 1992. Assessment and the improvement of education. *The curriculum journal*, 3(3):215–230.

Alon Jacovi and Yoav Goldberg. 2021. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C Wallace. 2020. Learning to faithfully rationalize by construction. *arXiv preprint arXiv:2005.00115*.

Xin Jiang, Hai Ye, Zhunchen Luo, WenHan Chao, and Wenjia Ma. 2018. Interpretable rationale augmented charge prediction system. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 146–151.

Klaus Krippendorff. 2004. Measuring the Reliability of Qualitative Text Analysis Data. *Quality and Quantity*, 38:787–800.

Himabindu Lakkaraju, Nino Arsov, and Osbert Bastani. 2020. Robust and stable black box explanations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5628–5638. PMLR.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

LightTag. 2018. The text annotation tool for teams. `https://www.lighttag.io/`.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Ian E Nielsen, Dimah Dera, Ghulam Rasool, Ravi P Ramachandran, and Nidhal Carla Bouaynaya. 2022. Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine*, 39(4):73–84.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566.

Ellis Batten Page. 1994. Computer grading of student prose, using modern concepts and software. *The Journal of experimental education*, 62(2):127–142.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. *arXiv preprint arXiv:2005.00652*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Julia Strout, Ye Zhang, and Raymond J Mooney. 2019. Do human rationales improve machine explanations? *arXiv preprint arXiv:1905.13714*.

Jana Z Sukkarieh, Stephen G Pulman, and Nicholas Raikes. 2003. Auto-marking: using computational linguistics to score short, free text responses.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Chul Sung, Tejas Dhamecha, Swarnadeep Saha, Tengfei Ma, Vinay Reddy, and Rishi Arora. 2019a. Pre-training bert on domain resources for short answer grading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6071–6075.

Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. 2019b. Improving short answer grading using transformer-based pre-training. In *International Conference on Artificial Intelligence in Education*, pages 469–481. Springer.

Daniel Vale, Ali El-Sharif, and Muhammed Ali. 2022. Explainable artificial intelligence (xai) post-hoc explainability methods: Risks and limitations in non-discrimination law. *AI and Ethics*, pages 1–12.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Dave Whittington and Helen Hunt. 1999. Approaches to the computerized assessment of free text responses.

Alistair Willis. 2015. Using nlp to support scalable assessment of short free text responses. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 243–253.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North*

American chapter of the association for computational linguistics; proceedings of the main conference, pages 260–267.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Ye Zhang, Iain Marshall, and Byron C Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 795. NIH Public Access.

Jiajie Zou, Yuran Zhang, Peiqing Jin, Cheng Luo, Xunyi Pan, and Nai Ding. 2021. Palrace: Reading comprehension dataset with human data and labeled rationales. *arXiv preprint arXiv:2106.12373*.

A Classification without fine-tuning

	Precision	**Recall**	**F1-score**
`Label 0`	0.49	0.46	0.47
`Label 1`	0.65	0.68	0.66
`Macro Avg`	0.57	0.57	0.57

Table 4: Overall classification performance metrics of SciBERT with no fine-tuning on question-answer pairs, where label 0 = Non-satisfactory and label 1 = Satisfactory.