International Journal of Assessment Tools in Education
2025, Vol. 12, No. 4, 905–925

https://doi.org/10.21449/ijate.1687429

journal homepage: https://dergipark.org.tr/en/pub/ijate *Research Article*

Answer-based and reference-based BERT models for automatic scoring of Turkish short answers: The decisive role of task complexity

Abdulkadir Kara [1*], Zeynep Avinç Kara [2], Serkan Yıldırım [3]

[1]Bayburt University, Department of Distance Education Application and Research Center, Bayburt, Türkiye
[2]TC Ministry of National Education, Erzurum, Türkiye
[3]Atatürk University, Kazım Karabekir Faculty of Education, Department of Computer Education and Instructional Technology, Erzurum, Türkiye

ARTICLE HISTORY

Received: Apr. 30, 2025

Accepted: Aug. 28, 2025

Keywords:

Artificial intelligence,
Deep learning,
Automatic scoring,
Short answer,
BERT LLM models.

Abstract: In measurement and evaluation processes, natural language responses are often avoided due to time, workload, and reliability concerns. However, the increasing popularity of automatic short-answer grading studies for natural language responses means such answers can now be measured more quickly and reliably. This study aims to build models for predicting automatic short answer scores using the pre-trained BERT deep learning language model and to reveal their effectiveness. For this purpose, two different score prediction models were created using an answer-based approach that aligns student answers with expert judgements and a reference-based approach that matches student answers with reference answers. The dataset includes answers from 246 Physics department students responding to 4 physics-related questions. The performance of these models was evaluated on four physics questions representing varying levels of cognitive complexity, using Cohen’s Kappa for statistical comparison of agreement with expert scores. Our findings reveal a clear interaction between model architecture and task complexity. The answer-based model was unequivocally superior for the most complex, multi-class task, effectively capturing diverse, nuanced responses. Conversely, the reference-based model demonstrated a statistically significant advantage for a well-defined, medium-complexity binary task. This study concludes that the optimal model for ASAG in Turkish is contingent on the cognitive demands of the assessment task, suggesting that a onesize-fits-all solution may not be the most effective approach. This provides a critical framework for practitioners, demonstrating not only that effective models are feasible for complex languages, but that their selection must be guided by task complexity.

1. INTRODUCTION

Assessing what and how students learn has become increasingly critical in today’s educational landscape, where instructional quality and accountability are closely intertwined. Measurement and evaluation are crucial in understanding educational effectiveness (Kurbanoğlu & Olcaytürk, 2023). Preferred understandings greatly influence students' learning outcomes (Yıldırım & Bilican-Demir, 2022). Choice-based techniques are commonly used in learning

*CONTACT: Abdulkadir KARA ✉ abdulkadirkara@bayburt.edu.tr Bayburt University, Department of Distance Education Application and Research Center, Bayburt, Türkiye

e-ISSN: 2148-7456

environments (Benli & İsmailova, 2018; Katsaris & Vidakis, 2021). Nevertheless, such techniques involve students selecting an option without justifying it (Çınar *et al.*, 2020), leading to random choices and undermining the validity of scores. Moreover, it can be difficult to identify lasting learning outcomes with this method. These issues arising from choice-based techniques require implementing measurement and evaluation techniques in academic settings and diversifying strategies.

Several factors stand out when considering the frequent use of choice-based techniques. The evaluation process of Natural Language Responses (NLR) increases the workload for teachers (Uyar & Büyükahıska, 2025). As a result, the evaluation process takes longer (Westera *et al.*, 2018). It becomes inapplicable, especially in crowded learning environments (Chen *et al.*, 2025). Additionally, there is a potential for including subjective evaluations from teachers within the score results, which may overshadow the process (Abdul-Salam *et al.*, 2022). Subjective judgements pose a risk to the reliability of scoring in evaluations. NLRs are less favored in learning environments due to increased workload, time requirements, and reliability concerns. Choice-based techniques are preferable as they offer quick and dependable measurements (Garg *et al.*, 2022; Hasanah *et al.*, 2016). However, they exhibit a limited ability to recognize learning situations. There is insufficient evidence for the detection of deep and meaningful learning with choice-based techniques (Noyes *et al.*, 2020). Students' ability to make random choices makes it difficult to accurately measure their cognitive levels (Zhu *et al.*, 2022).

The limitations imposed by choice-based techniques for identifying learning situations have increased research on Natural Language Processing (NLP) and NLR (Burrows *et al.*, 2015). Especially, technological developments have motivated research in this field (Jadidinejad & Mahmoudi, 2014). Some studies have addressed issues related to the structure of NLR to solve or mitigate the related problems. String-based research on NLRs has focused on word pairings and predicted sentence similarities (Leacock & Chodorow, 2003; Siddiqi *et al.*, 2010). Semantic-based research, which focuses on the meaning of the responses, uses pre-trained word vectors such as GloVe, FastText, and Word2Vec to analyze words semantically (Lubis *et al.*, 2021; Saunders *et al.*, 2014; Zehner *et al.*, 2016). With the advances in technology, machine learning and deep learning-based research has come to the forefront and is more effective in complex language structures than other research structures (Gomaa *et al.*, 2023; Li *et al.*, 2022; Tulu *et al.*, 2021; Uysal & Dogan, 2021). The reason for these efforts is that the advantages that the smoothly functioning NLR provides to the measurement and evaluation processes are important. Using NLR provides essential evidence for rigorous evaluation of the learning process (Westera *et al.*, 2018) and for developing and improving learning approaches (Noyes *et al.*, 2020). NLR provides and assesses learners' ability to accurately recall and communicate information (Uto & Uchida, 2020).

The common point of these studies in the literature is that they focus on automatic scoring of NLRs. Automatic scoring of NLRs stands out due to consistent and objective grading, reduced human labor, and time-saving rapid evaluation processes (Abdul-Salam *et al.*, 2022; Dönmez, 2024; Uyar & Büyükahıska, 2025). Automatic scoring studies on natural language began with Page's (1967) work as a secondary school teacher in the 1960s (Ramineni & Williamson, 2013). With the advancement in technology and research in natural language processing, its popularity has increased since the 2000s. Since 2010, numerous studies have been conducted on grading NLR (Filighera *et al.*, 2023; Ghavidel *et al.*, 2020; Saunders *et al.*, 2014; Tulu *et al.*, 2021; Zimmerman *et al.*, 2018). The rise of online learning environments has sparked interest in automatic scoring (Nath *et al.*, 2023). Especially the problems that emerged in the evaluation processes with the Covid-19 pandemic (Şenel & Şenel, 2021) can be considered to have an important share in drawing the attention of researchers to this field.

Burrows *et al*. (2015) identified seven types of NLR for automatic scoring: (1) fill-in-the-blank, (2) short answer, (3) essay, (4) structured text, (5) maths, (6) source code, and (7) voice and speech techniques. The objective of this study is to implement an automated system for the evaluation of short-answer questions. Short answers consist of only a few words or sentences (Nath *et al*., 2023). Burrows *et al*. (2015) distinguish short answers according to their length, focus, and clarity. In this answer type, the focus is on the meaning of the content. Short answers are objective and closed-ended in nature. The academic literature defines this domain as Automatic Short Answer Grading (ASAG). ASAG is a system that compares learner responses with one or more reference responses that are considered correct (Mohler & Mihalcea, 2009). Compared to manual systems, it can be said that these systems aim to provide justice more objectively (Badry *et al*., 2023).

When the literature is analyzed, it can be stated that research in the field of ASAG has increased in recent years. Noticeably, English NLR datasets are used more frequently in the studies. However, developing pre-trained language models that can be used for many languages has facilitated research on different languages. Of particular interest are large language models (LLMs) such as BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-Trained Transformer), XLNet (Extra Long Network) and RoBERTa (Robustly optimised BERT approach), which also allow studies in multiple languages (Abdul-Mageed *et al*., 2020; Zhang & Copus, 2023). This is because pre-trained models such as BERT can perform well on specialized tasks such as automatic scoring by leveraging large datasets to improve accuracy and efficiency (Chan *et al*., 2024). The impact of LLMs is evident in the positive results observed in recent ASAG research on different languages (Mardini *et al*., 2024; Sawatzki *et al*., 2021; Sung *et al*., 2019). These developments are promising for the dissemination of ASAG in different languages. This study aims to develop and evaluate two BERT-based ASAG models—an answer-based and a reference-based model—for the automatic scoring of Turkish short answers.

1.1. Related Work

The history of automatic scoring studies dates back to the 1960s (Page, 1967). With the developments in the field of NLP, interest in the ASAG field has recently increased. Different technological components have been utilized in the studies conducted in the historical process. Examples include word matching algorithms, similarity measures focusing on semantic distance, vector space representations, and machine learning algorithms. Burrows *et al*. (2015) classified the studies in the field of ASAG according to model approaches. The approaches are concept mapping, knowledge extraction, corpus-based, and machine learning. In another study, ASAG approaches are considered as similarity-oriented: (1) string-based, (2) semantic-based, (3) hybrid-based, and (4) machine and deep learning (Abdul-Salam *et al*., 2022). It can be pointed out that attention is paid to historical development processes in approach classifications. In recent years, it is noteworthy that ASAG models have been mainly developed with a deep learning approach (Chaudhari & Patel, 2024). Research combining LSTM and derivative models with large language models (LLM), such as BERT, has achieved successful results (Gomaa *et al*., 2023; Li *et al*., 2022).

In line with the focus of the research, some end-use applications using the pre-trained BERT model were examined in the literature. We focused on BERT studies on different languages in the literature and general ASAG studies on Turkish. Zhu *et al*. (2022) developed a pre-trained BERT-based neural network model for the ASAG system. Their research included a semantic refinement layer consisting of Bi-LSTM (Bidirectional Long Short-Term Memory) and Capsule networks to improve the meaning of BERT outputs. The findings indicate that the developed model successfully obtained favorable outcomes compared to most other techniques and methods across the SemEval-2013 and Mohler datasets. In another study on SemEval-2013 tasks, BERT and XLNET pre-trained deep learning models were applied to the ASAG system

(Ghavidel *et al*., 2020). In both models, improved results were achieved compared to previous studies. Sung *et al*. (2019) concentrated on advancing the existing BERT model by incorporating texts from related subject areas. They also created a second model through fine-tuning, which considered the student and reference answer pair for additional answer prediction. They designed a dataset of 3 sub-topics in the industrial field. The study's experimental results demonstrate that the BERT model, developed with text-based content during its training phase, outperformed the fine-tuned model. Amur *et al*. (2022) applied a BERT-based ASAG application with sQuad 2.0 dataset to students in Roshan Tara school in Mehrabpur, Pakistan. The score predictions of the system were found to be highly successful.

Mardini *et al*. (2024) developed a dataset in Spanish targeting university students across ten subject areas, comprising 3772 answers scored from 0 (incorrect) to 5 (correct) and from 0 (incorrect) to 1 (correct). The study constructed the BERT model in six distinct approaches, and the answers were evaluated in both English and Spanish. In addition, the Skip-thought approach was also employed in the study. The study results indicated the potential contribution of the fine-tuned BERT model in developing reading comprehension skills in various languages. Nath *et al*. (2023) conducted experiments using the CREG (Meurers *et al*., 2011) and CSSAG (Padó, 2016) datasets in German to develop an ASAG system. The BERT model yielded better results compared to the similarity-based approach utilizing bag-of-words. Notably, the CREG dataset provided more effective outcomes. Sawatzki *et al*. (2021) achieved highly successful outcomes compared to previous studies by employing a similar BERT approach for German and English datasets. The study used feature extraction architecture and fine-tuning with datasets from the Business Administration undergraduate program and the University of North Texas.

Nael *et al*. (2022) conducted one of the initial BERT studies on the Arabic language. They employed the Arabic adaptation of the Kaggle ASAP dataset to produce an ASAG system based on deep learning. Their study compared two novel approach models, BERT and ELECTRA (Efficiently Learning an Encoder That Accurately Classifies Token Replacements), with conventional deep learning models. The outcomes indicated that applying new approaches led to better results for short-answer scoring models. Schleifer *et al*. (2023) conducted a comparative study to ascertain superior performance. The AlephBERT PLM (Seker *et al*., 2022) was employed to scrutinise Hebrew and the ASAG system was created with a dataset covering Biology topics. A performance comparison was, then, conducted between the AlephBERT-based and Convolutional Neural Network (CNN)-based systems. The study revealed that the AlephBERT-based model outperformed the CNN-based model.

Studies in the literature show that it is possible to develop ASAG systems with increased efficiency and reduced workload in various languages. Creating ASAG systems using traditional models typically demands in-depth feature engineering and considerable fine-tuning (Salim *et al*., 2022). The chance to create more effective systems without lengthy NLP processes has prompted researchers to turn to pre-trained models like BERT (Chen *et al*., 2023; Haller *et al*., 2022). Analysis of studies indicates that BERT deep learning models have gained widespread adoption in the ASAG field recently. The data sets used in the systems developed with the BERT model and the results obtained are summarized in Table 1.

When examining the Turkish ASAG literature, it becomes apparent that only a limited number of studies have been conducted. To address this gap, Çınar *et al*. (2020) sought to develop a machine learning-based system to score Turkish short answers automatically. They collected a dataset containing answers from university-level physics students. SVM (Support Vector Machines), Gini, KNN (k-Nearest Neighbours), Bagging and Boosting techniques were used for model development. The highest model performance was obtained with AdaBoost.M1. This study drew attention as one of the pioneering studies in Turkish ASAG. The study conducted by Uysal and Doğan (2021) also consisted of short-answer items. The dataset comprised limited open-ended answers in the field of Turkish from the ABIDE program conducted by the Ministry of National Education. Restricted open-ended items can be seen as equivalent to short-answer items. In their study, deep learning models were also used along with machine learning. The study evaluated the consistency between the automatic scoring mechanism created in the study and the scores provided by experts. The automatic scoring system development process employed five algorithms: SVM, LR (Logistic Regression), MNB (Multinomial Naive Bayes), LSTM, and BLSTM. Successful findings were recorded in this study, encouraging further investigation into automatic scoring studies on Turkish.

Table 1. *Performance results of ASAG models developed with BERT.*

Author	Dataset	Results
Sung *et al*. (2019)	SemEval-2013	Accuracy = .759, F1 = .758
Ghavidel *et al*. (2020)	SemEval-2013	Accuracy = .798, F1 = .797
Sawatzki *et al*. (2021)	German dataset	Quadratic weighted kappa (QWK) = .82, *r* = .892
Amur *et al*. (2022)	SQuad 2.0	QWK = .77, F1 = .96, Precision = .95
Nael *et al*. (2022)	ASAP-SAS	QWK = .77
Zhu *et al*. (2022)	SemEval-2013&Mohler	QWK = .82, *r* = .892
Schleifer *et al*. (2023)	Biology dataset	QWK ≥ .90
Mardini *et al*. (2024)	Spanish dataset	RMSE = .59, *r* = .78

The available literature on Turkish ASAG studies indicates their limited number. Upon analysis of the existing studies, it is evident that traditional machine learning and deep learning models are prominent in the system development process. Distinct from these studies, our research aims to develop the Turkish ASAG system using BERT, a popular pre-trained deep learning model in the field. In this context, our investigation represents one of the pioneering endeavors in establishing a Turkish ASAG system utilizing BERT.

When the literature is examined, it is seen that answer-based and reference-based approaches stand out in the BERT model development process. In the answer-based approach, the training process is carried out by establishing a relationship between the Student Answer (SA) and the Expert Scores (ES) defined for all SAs. In the reference-based approach, the training process is carried out by establishing a similarity relationship between the SA and the Reference Answer (RA) (Nael *et al*., 2022). Both approaches were considered in our study, and the automatic scoring of short Turkish answers was emphasized. This research constituted one of the pioneering studies in the field of automatic scoring of short answers in Turkish (ASAG) based on the BERT deep learning model, which has shown great success in recent years. The fact that BERT-based ASAG applications in Turkish have not yet been sufficiently covered in the literature increases the potential of this study to fill an important gap in the field and to serve as a basis for future research.

This study aims to develop and evaluate two BERT-based ASAG models—an answer-based and a reference-based model—for the automatic scoring of Turkish short answers. The research questions determined for the study in line with the research purpose are as follows;

- What is the scoring performance of BERT-based, answer-based and reference-based models in automatically evaluating Turkish short answers?
- How does the cognitive complexity of the assessment task (e.g., binary vs. multi-class, lower vs. higher-order thinking) influence the comparative performance of these models?

2. METHOD

Concurrently with ongoing research in this field, we incorporated the pre-existing BERT deep learning model in developing a Turkish ASAG system. The study utilized a comparative experimental research approach to evaluate two training methods for the BERT deep learning model to improve a Turkish ASAG system. Consequently, the results obtained from the developed models were analyzed and interpreted to determine which method was optimal for development.

2.1. Dataset

In the study, the Physics data set was developed by Çınar *et al*. (2020). Two hundred forty-six (246) students studying in the Physics department at a state university responded to 4 physics-related questions. Inter-rater reliability was ensured by providing the dataset format for the answers evaluated by three experts (Çınar *et al*., 2020). Inter-rater reliability values were calculated by the Pearson Correlation Coefficient. The correlations between the scores given by the raters to the short answer questions were .87 for question 1, .79 for question 2, .92 for question 3, .90 for question 4 and .87 for the average correlation. The average reliability correlation value of all questions was calculated by averaging the correlation values of the questions. The dataset contained student answers, reference answers, and expert scores. The reference answers were created as scoring keys. The question items and scoring keys in the data set were presented clearly in Appendix 1. The study of Çınar *et al*. (2020) provided more detailed information about the question items.

Answers for Q1, Q3, and Q4 were objectively evaluated on a binary scale, while answers for Q2 received scores ranging from 0 to 4. The characteristics of the questions in the data set are presented in Table 2, while Table 3 shows the base data set properties.

Table 2. *Question characteristics.*

Question ID	Topic	Bloom's Taxonomy	Scoring Type
Q1	Electricity	Comprehension level	0-1
Q2	Conservation of energy	Comprehension level	0-1-2-3-4
Q3	Energy	Knowledge stage	0-1
Q4	Work	Knowledge stage	0-1

When the content of the dataset was analyzed, it was observed that it focused on core topics in physics, including electricity, energy, and work. The question structures, when classified according to Bloom's Taxonomy, aligned with the lower cognitive levels—specifically, knowledge and comprehension. However, despite all questions falling under these categories, they exhibited varying degrees of cognitive complexity in terms of the type of reasoning and elaboration required. For example, while Q1, Q3, and Q4 required relatively straightforward recall or identification of concepts (and were scored using binary scoring: 0 or 1), Q2 was designed to assess a more nuanced understanding and explanation of scientific reasoning, and thus used a multi-level scoring system (0–4). This task-based variability enabled us to examine how model performance was affected by differences in scoring structure and cognitive complexity, even within the same taxonomy level.

Table 3. *Base Physics dataset properties.*

Question	Answer Number	Distribution of Scores*	Scoring Type
Q1	254	89 / 165	0-1
Q2	147	73 / 6 / 8 / 24 / 36	0-1-2-3-4
Q3	254	31 / 223	0-1
Q4	254	155 / 99	0-1

*The distribution of scores for Q1, Q3, and Q4 indicates the number of labels 0 and 1, and for Q2 indicates the number of labels 0, 1, 2, 3, and 4, respectively.

When the dataset was analyzed in detail in Table 3, it was seen that the class distributions were significantly imbalanced. In the Distribution of Scores column, the scores given to the answers were presented in order. For Q1, there were 165 correct answers, while the number of incorrect answers was 89. In Q3, the number of incorrect answers was 31, while the number of correct answers was 223. In Q2, it was observed that the number of collective responses needed to be higher, and the Distribution of the number of responses from the classes needed to be more balanced since multiple scoring type was used. Q4 was found to have a relatively more balanced scoring distribution than the others. For Q4, the number of incorrect answers was 155 while the number of correct answers was 99.

Using this dataset for the first time, Çınar *et al*. (2020) found a solution to the imbalanced dataset problem by using the Smote (Synthetic Minority Over-sampling) over-sampling technique. In this study, generative artificial intelligence (GenAI) technologies were used to balance the classification distributions in the dataset. The problem of class imbalance in the dataset was based on a sampling approach with equal representation of subgroups proposed by Zhang *et al*. (2024). As a result, the derived Physics dataset was ready for application in the study and the class imbalance problem had been solved. Table 4 illustrates the properties of the updated dataset.

Table 4. *Updated Physics dataset properties.*

Question	Answer Number	Distribution of Scores*	Scoring Type
Q1	265	100 / 165	0-1
Q2	266	66 / 50 / 50 / 50 / 50	0-1-2-3-4
Q3	274	100 / 174	0-1
Q4	254	150 / 128	0-1

*The distribution of scores for Q1, Q3, and Q4 indicates the number of labels 0 and 1, and for Q2 indicates the number of labels 0, 1, 2, 3, and 4, respectively.

Empty answers in the Base Dataset were removed, and new answers that resembled student answers were derived, especially to balance the class distribution. The dataset had undergone minor modifications. The objectives of these actions were to enable the developed models to categorize the classes accurately and to equilibrate the class distribution in data classification partially. No data duplication via repetition was performed during data derivation. The purpose of these procedures was to ensure that the class distribution in the data set was balanced.

2.2. Model Design

In the study, the “dbmdz/bert-base-turkish-uncased” model shared on the HuggingFace platform was preferred for automatic scoring of Turkish short answers. This model is a BERT model trained on large-scale text corpora, considering the specific features of Turkish language structure, and is considered a base model with high performance in Turkish natural language processing tasks. The “dbmdz/bert-base-turkish-uncased” model follows the original BERT architecture (Devlin *et al*., 2019), including 12 transformer layers, 768-dimensional hidden layers, and 12 attention heads. The model contains approximately 110 million parameters and is pre-trained on 32GB of Turkish text.

The first method used student answers (SA) and expert scores (ES) for model training. The second method used the similarity between SA and reference answers (RA) to train the model. This study resulted in the development of two models suitable for both answer-based and reference-based training approaches. Figures 1 and 2 depict the general structure of the models designed for the automatic scoring of concise answers. The structures of both models are remarkably similar. They comprise the dataset and model processes indicated in Figures 1 and 2. Model 1, developed through answer-based training, has been created using expert scores. Model 2, developed through reference-based training, employs a similarity-based approach to system predictions. The dataset processing of both models followed similar stages. The dataset was initially selected and partitioned into training (70%), validation (15%), and test (15%) sets for Model 1. For Model 2, the same partitioning approach was applied with training (70%), validation (15%), and test (15%) subsets. BertTokenizer, a large language model created for the Turkish language, was implemented in the tokenization process for the models. Therefore, the text data was transformed into a format appropriate for the BERT model. Model 1 used the Classification (CLS) embedding vector to derive features from all processed text inputs. CLS was particularly noteworthy in the realm of text classification (Devlin *et al*., 2019; Sun *et al*., 2019). This element contributed to the effectiveness of considerable bidirectional language models like BERT (Yang *et al*., 2022).

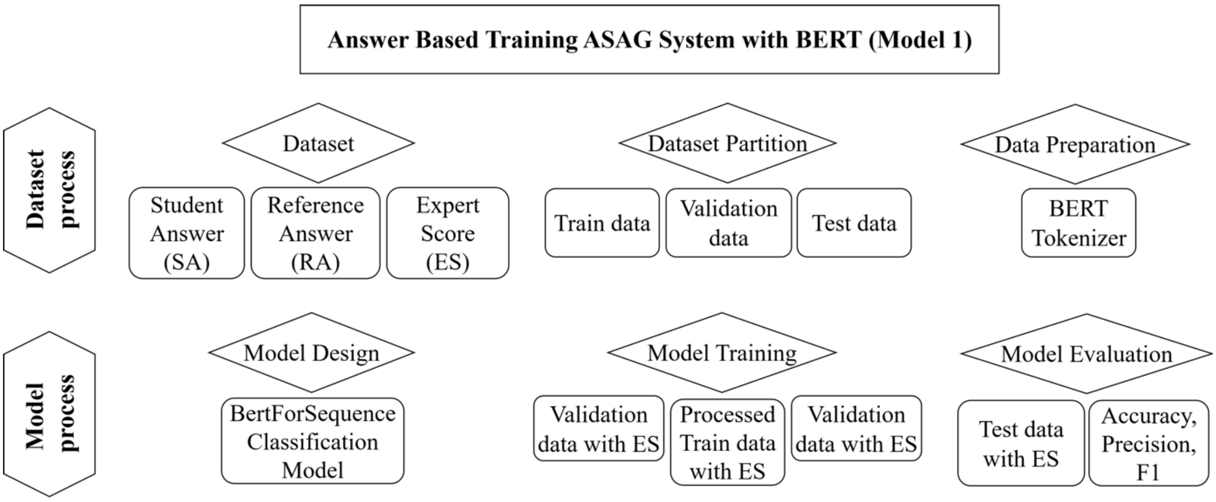

Figure 1. *Answer-based training ASAG system with BERT (Model 1).*

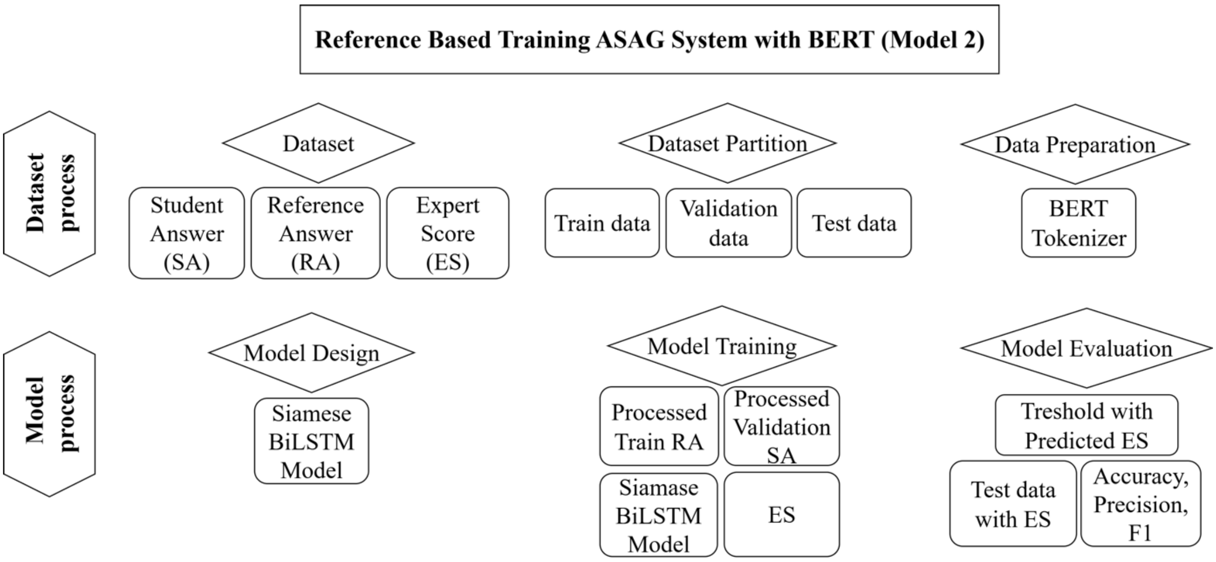

Figure 2. *Reference-based training ASAG system with BERT (Model 2).*

The value information for the basic parameters we used in the model development process is presented in Table 5.

Table 5. *Parameter values*

Parameter	Value
Max length	128
Learning rate	2e−5
Batch size	32
Epochs	10
Early stopping patience	3
Dropout probability	.1

During the modeling process, the two approaches diverged in their training methods. Model 1 utilized the "BertForSequenceClassification" class to categorize text data for the Turkish BERT model. Training directly aligned expert scores with student answers in the processed training set. In Model 2, the development process of the BERT model was executed utilizing Siamese BiLSTM. The model produced a similarity score between the answers of the students and those of the reference. The loss between the acquired similarity scores and expert scores was calculated to update the model. Adam optimizer and an epoch size of 10 were selected for optimization.

With large data sets, developing models with high-accuracy performance without overfitting problems was more practical. We aimed to overcome the disadvantage caused by the small size of our dataset with as simple a model setup as possible. We divided our dataset into training, testing, and validation at random intervals. Thus, we could directly monitor validation losses during training to avoid overfitting the models. The model training was terminated with the early stopping function when the performance stabilized. In our study, we also included the dropout technique to reduce the overfitting tendencies of the models. The dropout technique allowed model training to occur in multiple ways. Because some of the neurons were randomly deactivated in each training phase. The use of dropout strengthened the flexibility and generalizability of our models.

2.3. Data Analysis

The model development process and performance analysis of the developed models were carried out through Google Colab. The runtime type was “python 3”. “A100 GPU” was preferred for virtual GPU usage. Various evaluation metrics were applied to determine the effectiveness of ASAG models (Zesch *et al*., 2023). In research studies, metrics such as accuracy, agreement, and correlation are commonly employed (Burrows *et al*., 2015). Classification models commonly measure their performance using accuracy, precision, and F1 scores (Chaudhari & Patel, 2024). Accuracy is the ratio of correct model predictions to total answers, while precision is concerned with the number of true and false positives. The F1 score is the harmonic mean of precision and recall values. It is generally preferable to see a better model performance in cases with unbalanced class distributions (Riyanto *et al*., 2023). These metrics were used to evaluate the performance of the developed models.

In addition to these individual performance metrics, a direct statistical comparison between Model 1 and Model 2 was conducted to identify significant differences in their predictions. For this purpose, a dual-analysis approach was adopted. First, McNemar's test was employed. This non-parametric test is specifically designed for paired nominal data and is used to determine if the two models have differing error rates. It evaluates whether one model is significantly more likely to be correct when the other is incorrect. Second, Cohen's Kappa (κ) was calculated as a measure of inter-rater agreement between each model and the expert scores. Unlike simple accuracy, Cohen's Kappa accounts for the possibility of agreement occurring by chance, providing a more robust measure of performance, especially in multi-class or imbalanced datasets. All statistical analyses were performed using the Scikit-learn library in Python.

The data analysis also included a qualitative error analysis to complement the quantitative findings. This analysis focused on instances where the models' predictions diverged from the expert scores, particularly on the more complex, multi-class task (Q2). The objective of this analysis was to identify systematic error patterns, understand the types of student answers that would pose challenges for the models, and gain deeper insights into the models' decision-making processes. The process involved a manual review of misclassified responses to categorize the nature of the errors. The implementation of all statistical and qualitative analyses relied on the Scikit-learn and Pandas libraries in Python.

3. RESULTS

Model 1 was the ASAG model, developed through an answer-based learning approach. The performance of each request within the dataset was extracted and measured using accuracy, precision, and F1 score. The corresponding results for Model 1's performance are presented in Table 6.

Table 6. *Answer-based (Model 1) ASAG performance.*

Question	Accuracy	Precision*	F1 Score**
Q1	.82	.82 / .82	.82 / .82
Q2	.90	.91 / .91	.90 / .90
Q3	.94	.96 / .92	.93 / .94
Q4	.86	.86 / .86	.86 / .86

*Precision values represent both macro and weighted values.
**F1 Score values also represent both macro and weighted values.

Upon analyzing the performance results of Model 1, it can be concluded that the model successfully dealt with the physics dataset obtained. Accuracy scores ranging from .82-.94 were achieved for all questions by the Accuracy metric. The Precision results were also promising, with macro scores of .82-.96 and weighted scores of .82-.92. While macro average values for F1 scores were .82-.93, weighted values were .82-.94. The main difference between macro and weighted average is that the macro average gives equal importance to each class, while the weighted average takes into account class imbalance. The macro average calculates and averages the metric for each class independently. Weighted average calculates the metric for each class but weights it according to the number of samples in that class. In the context of this study, the macro precision/F1 score gave equal importance to each scoring category, while the weighted precision/F1 score gave more importance to scoring categories with more student responses.

Model 2 represents the ASAG model, developed using a reference-based learning approach. The prediction results were assessed utilizing accuracy, precision, and F1 scores, much like in Model 1. The performance data of Model 2 are presented in Table 7.

Table 7. *Reference-based (Model 2) ASAG performance.*

Question	Accuracy	Precision*	F1 Score**
Q1	.85	.90 / .88	.83 / .84
Q2	.62	.64 / .62	.61 / .57
Q3	.95	.94 / .95	.94 / .95
Q4	.86	.86 / .86	.86 / .86

*Precision values represent both macro and weighted values.
**F1 Score values also represent both macro and weighted values.

The results of Model 2 also yielded a satisfactory outcome for this dataset. Accuracy scores ranged from .62 to .95. Precision macro values ranged from .64 to .94, while weighted average values ranged from .62 to .95. F1 scores for macro were between .61 and .94, while weighted average values ranged from .57 to .95. As shown in Table 7, the accuracy and precision values were in harmony. Thus, the accuracy performance of Model 2 can be generalized to the new data and is not subject to overfitting.

The performance metrics for Model 2, the reference-based ASAG model, are presented in Table 7. The model's performance varied significantly across the different tasks. Accuracy scores ranged from a low of .62 to a high of .95. A similar spread was observed in the F1 scores, with weighted F1 scores spanning from .57 to .95. Notably, the model's lowest performance metrics were consistently recorded on the multi-class scoring task (Q2), with an accuracy of .62 and a weighted F1 score of .57.

A preliminary review of Tables 6 and 7 indicates that both models perform effectively on the dataset, but a more direct and statistically robust comparison is necessary to discern significant performance differences. To achieve this, a formal statistical comparison was conducted using a dual-analysis approach. First, McNemar's test was employed to identify statistically significant differences in the error rates of the two models. Second, Cohen's Kappa (κ) was calculated to provide a more nuanced measure of agreement between each model's predictions and the expert scores, which is particularly suitable for tasks with multiple or ordinal scoring categories. The comprehensive results of this comparative analysis are presented in Table 8.

Table 8. *Comprehensive comparison of model performance.*

Question	Task Type	McNemar *p*-value	Superior Model	Model 1 Kappa (κ_1)	Model 2 Kappa (κ_2)
Q1	Binary	< .001	Model 2	.575	.681
Q2	Multiple	< .001	Model 1	.843	.532
Q3	Binary	> .999	n.s	.952	.952
Q4	Binary	> .999	n.s*	.631	.678

*n.s. = not significant. While the Kappa scores for Q4 show a numerical difference, the McNemar test indicates that this difference is not statistically significant.

For the multi-class scoring task (Q2), a statistically significant difference in performance was identified (McNemar's test, $p < .001$). Model 1 demonstrated a substantially higher agreement with expert scores ($\kappa_1 = .843$) compared to Model 2 ($\kappa_2 = .532$). A significant difference was also found for question Q1 (McNemar's test, $p <. 001$); however, in this case, Model 2 achieved a higher agreement score ($\kappa_2 = .681$) than Model 1 ($\kappa_1 = .575$). Finally, for the binary classification tasks Q3 and Q4, no statistically significant difference was detected between the models ($p > .999$). Their Cohen's Kappa scores were also highly comparable for these questions ($\kappa_1 = .952$ vs. $\kappa_2 = .952$ for Q3, and $\kappa_1 = .631$ vs. $\kappa_2 = .678$ for Q4).

To further investigate the quantitative performance of Model 2, particularly its low agreement score ($\kappa = .532$) on the multi-class task (Q2), a qualitative error analysis was conducted on its predictions. The analysis revealed several systematic error patterns. The most prominent error pattern was the systematic misclassification of answers with a true score of '2', which were nearly all incorrectly assigned a score of '0'. These misclassified responses were typically characterized by their conciseness. While they correctly stated the core scientific principle, it was observed that their presentation was very direct. Examples of such answers include:

- *Hızları değişmez hız kütleye bağlı değildir [The speeds do not change; speed is not dependent on mass]*.
- *Kütlenin düşme hızı bağlamında bir etkisi yoktur, aynı hızda düşerler [Mass has no effect in the context of falling speed; they fall at the same speed]*.
- *Aynıdır serbest düşme mantığına göre kütlenin bir önemi yoktur [It is the same; according to the logic of free fall, mass is not important]*.

A common feature of these answers is their lack of detailed explanations involving concepts such as potential or kinetic energy, which are present in higher-scoring responses and the reference answer.

A second observed error pattern is the model's inconsistent differentiation between scores of '3' and '4'. The model occasionally assigns a different score to answers that are textually and conceptually very similar. Finally, a notable outlier is an instance where the model assigns a score of '0' to a detailed, conceptually rich answer with a true score of '3'. This particular answer introduces a related concept ("air resistance") not present in the reference answer.

4. DISCUSSION and CONCLUSION

This study developed and compared two BERT-based models for the automatic scoring of Turkish short answers: an answer-based (Model 1) and a reference-based (Model 2) model. Our results revealed that the choice of the optimal scoring model was not absolute but was contingent on the cognitive complexity of the assessment task.

4.1. The Impact of Task Complexity on Model Performance

The central results of this research are the clear interaction between model architecture and task complexity, which manifested across a spectrum of assessment types. For the most cognitively demanding, multi-class task (Q2), the answer-based model (Model 1) was unequivocally superior, with a significantly higher agreement score than its counterpart ($\kappa_1 = .843$ vs. $\kappa_2 = .532$, $p < .001$). This suggests that when answers require nuanced understanding and can be expressed in many valid ways, a model trained on a diverse set of student responses is better equipped to learn the complex patterns of partial and full credit. Conversely, for the medium-complexity task of identifying a specific, well-defined misconception (Q1), the reference-based model (Model 2) demonstrated a statistically significant advantage ($\kappa_2 = .681$ vs. $\kappa_1 = .575$, $p < .001$). Similarly, Sayeed and Gupta (2022) emphasized that reference-based approaches demonstrated superior performance, particularly for medium and lower complexity tasks, achieving significant improvements in ASAG systems when utilizing Siamese-based transformers that model the evaluation as sentence similarity between reference and student answer pairs. Finally, for the low-complexity definitional tasks (Q3 & Q4), both models performed at a high level with no statistically significant difference, suggesting the architectural choice was less critical for simple knowledge recall. These results further validated the observations by Zhu *et al*. (2022), who emphasized that fine-tuned BERT models could achieve successful results even with small corpora, demonstrating the robustness of transformer-based architectures across varying dataset sizes and task complexities. Similarly, Salim *et al*. (2022) emphasized that pre-trained BERT models, through their transfer learning capabilities and effective fine-tuning processes, could achieve high performance even with limited data.

4.2. Understanding Model Limitations: A Qualitative Perspective

The reference-based model (Model 2), while effective in binary classification tasks, exhibited a significant limitation in the multi-class scoring task, particularly in its complete inability to identify answers corresponding to the intermediate 'Score 2' category. Qualitative error analysis suggests this issue stems from the model's fundamental design, which equates semantic similarity to a single, high-quality reference answer with scoring accuracy. This approach systematically penalized answers that were conceptually correct but concise, as they lacked the detailed phrasing and specific keywords present in the comprehensive reference answer. Intermediate scores may have become an ambiguous 'middle ground' that the model struggled to learn. In this situation, the model's struggle to differentiate between a student's grasp of the core concept and their ability to articulate it in a textually similar manner may stem from a key vulnerability inherent in single-reference-based approaches for nuanced, multi-level assessment scenarios. A potential solution to this problem is to integrate multiple reference answers representing diverse yet pedagogically valid expressions. This may help the model capture a broader semantic space and reduce the penalization of concise but correct answers. Similarly, Akila-Devi *et al*. (2023) emphasized that reference-based approaches significantly enhanced performance in ASAG systems, particularly when reference answers were strengthened through diverse content acquisition from multiple sources, including expert responses and community question-answering platforms. Indeed, implementing a strategy such as clustering student responses thematically and using centroid-based representations as auxiliary references could support better outcomes, improving the model’s sensitivity to variation in student phrasing, especially in mid-range scoring categories.

The challenges we observed with our reference-based model's inability to handle intermediate scoring categories could be attributed to dataset imbalance or insufficient data issues, particularly affecting multi-class classification tasks that required nuanced scoring distinctions. Similarly, Mardini *et al*. (2024) emphasized in the literature that reference-based approaches showed low performance in predicting extreme values due to imbalanced data distribution combined with insufficient data in multi-class classification tasks. In addition, employing semi-supervised learning techniques—such as pseudo-labeling or confidence-based refinement—could help the model learn more effectively from unlabeled or uncertain responses. This approach allowed the model to generate provisional labels for ambiguous student answers and used them in further training cycles. Specifically, it could help the model handle mid-range scores more accurately, which were often challenging not only due to underrepresentation but also because such answers tended to be concise or expressed in diverse, non-standard ways that deviated from the reference answer's phrasing. By incorporating these borderline or confidently predicted samples, the model can better be generalized across varying answer styles and levels of elaboration. Similarly, Xie *et al*. (2023) emphasized that pseudo-labeling enhanced class prediction accuracy, as demonstrated by their CAP (Class-Aware Pseudo-Labeling) method.

4.3. Contextualizing Performance Within the Field

The performance metrics achieved in our study are highly competitive, aligning closely with results from BERT-based ASAG systems in other languages such as German (Sawatzki *et al*., 2021), Spanish (Mardini *et al*., 2024) and Arabic (Nael *et al*., 2022). This demonstrates that the effectiveness of transformer-based architectures for automatic scoring is not language-specific and that Turkish, as a morphologically rich language, can similarly benefit from these advanced models.

The advantages of cross-linguistic consistency, efficiency and generalizability became even clearer when comparing our results to the work of Çınar *et al*. (2020), who used the same dataset. M1 model achieved a marginally higher peak F1 score; our BERT models reached a comparable level of performance without the need for extensive, manual feature engineering. This distinction is critical, as traditional machine-learning approaches are known to be complex and require significant manual intervention (Burrows *et al*., 2015; Zehner *et al*., 2016). Therefore, the key advantage of our approach lies in its efficiency. By leveraging pre-trained models, we demonstrate a more direct and scalable pathway for developing high-performing ASAG systems. This result resonates with the broader trend in the field away from feature-dependent models (Abdul-Salam *et al*., 2022) and can provide strong motivation for the continued development of BERT-based models for Turkish.

4.4. Limitations

It is important to acknowledge several limitations of this study, most of which pertain to the dataset and its scope. These limitations may influence the generalizability and interpretability of the findings, particularly in relation to model behavior across different tasks and contexts.

- Domain Limitation: The dataset was limited to a single subject domain—Physics—which may restrict the applicability of the results to other disciplines with different linguistic or conceptual characteristics.
- Cognitive Level Limitation: All questions in the dataset targeted only the knowledge and comprehension levels of Bloom's Taxonomy. Therefore, the findings may not extend to higher-order cognitive tasks such as application, analysis, or evaluation, where students' language use and reasoning strategies may differ substantially.
- Data Size and Diversity: The study involved a relatively small dataset, including only four questions and responses from 246 students. This modest scale may not capture the full variability present in real-world educational settings. In particular, multi-class classification tasks may require more data to accurately model nuanced score boundaries, as noted in recent literature (Mardini *et al*., 2024).
- Language-Specific Bias: This study's exclusive focus was on Turkish—a morphologically rich and agglutinative language—may limit the generalizability of our findings to languages with different linguistic structures. Nonetheless, the strong performance of our models suggests that transformer-based architectures like BERT could effectively handle such

complexity. While this remains a limitation, it also positions Turkish as a promising test case that may encourage further research across diverse language families.

Given these limitations, caution should be exercised when interpreting the results beyond the context of this specific dataset. As the reliability and validity of models are best understood across different datasets (Zhu *et al*., 2022), future work should aim to test the durability and sustainability of these models on a wider range of subjects and different cognitive tasks.

4.5. Implications

In summary, this research demonstrates the effectiveness of BERT-based models for the automatic scoring of short answers in Turkish. An important further contribution of the study is the result that the optimal model architecture is not universal but is contingent on the cognitive complexity of the assessment task. Specifically, an answer-based approach excels for complex, multi-faceted questions, while a reference-based model is more reliable for identifying specific, well-defined concepts.

Furthermore, this research establishes that these pre-trained models can achieve performance levels comparable to traditional machine learning techniques while significantly reducing the need for laborious feature engineering. This can provide a powerful and potentially more efficient pathway for developing ASAG systems for Turkish and make a significant contribution to a field where such applications are limited. While the promising performance indicates significant potential for practical applications, the generalizability of these findings requires further research using larger and more diverse datasets.

4.6. Recommendations and Future Work

4.6.1. *Model-Specific Recommendations*

In light of the results obtained from this study—particularly the limitations observed in the reference-based model during multi-class scoring—several potential improvements can be considered to enhance model performance in similar contexts:

- Incorporating multiple reference answers may help the model better recognize semantically correct but textually diverse student responses, especially for intermediate score categories. By capturing a broader range of acceptable expressions, this approach could reduce the likelihood of penalizing valid but concise answers.
- Thematic clustering of student responses and the use of centroid-based representations as auxiliary references might contribute to improving the model's robustness against lexical and structural variation. This data-driven approach could complement manually created references by reflecting how students naturally express their understanding.
- Semi-supervised learning methods, such as pseudo-labeling or confidence-based refinement, may offer a way to utilize ambiguous or unlabeled responses more effectively. These techniques could support the model in learning from borderline or underrepresented cases and improve its ability to generalize in nuanced scoring tasks.

These strategies do not represent definitive solutions, but they may provide useful directions for addressing the challenges encountered with reference-based ASAG models in future work.

4.6.2. *Broader Research Directions*

Beyond the technical aspects of the models, several broader areas of inquiry may also support the advancement of ASAG systems for Turkish and similar languages:

- New insights can be brought to the field by implementing the ASAG models in learning environments and taking the opinions of teachers and students regarding performance status.
- Different LLM models developed for Turkish can be compared with the performance of ASAG, and even more reliable performance outputs can be obtained by considering these models together.

- In the future, using automatic formative feedback systems with ASAG applications will strengthen the communication between students and educators in the follow-up of learning situations.
- The data set used in this study corresponds to only two thinking processes in Bloom's Taxonomy: knowledge and comprehension. In future studies, ASAG performances can be investigated on data sets for higher cognitive levels in Bloom's Taxonomy.
- Additional research may also examine ways to enhance the explainability of automatic scoring decisions, thereby increasing transparency and trust among educators and learners.

Acknowledgments

We thank Çınar *et al*. (2020) for sharing the physics dataset as an open data source.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Contribution of Authors

Abdulkadir Kara: Investigation, Resources, Methodology, Software, Formal analysis, and Writing-original draft. **Zeynep Avinç Kara:** Investigation, Resources, Validation, and Formal analysis. **Serkan Yıldırım:** Methodology, Supervision, Validation, and Writing-original draft.

Orcid

Abdulkadir Kara https://orcid.org/0000-0003-3255-1408
Zeynep Avinç Kara https://orcid.org/0000-0002-8309-3876
Serkan Yıldırım https://orcid.org/0000-0002-8277-5963

REFERENCES

Abdul-Mageed, M., Elmadany, A., & Nagoudi, E.M.B. (2020). *ARBERT & MARBERT: Deep bidirectional transformers for Arabic*. arXiv. https://doi.org/10.48550/arXiv.2101.01785

Abdul-Salam, M., El-Fatah, M.A., & Hassan, N.F. (2022). Automatic grading for Arabic short answer questions using optimized deep learning model. *PloS ONE*, *17*(8), Article e0272269. https://doi.org/10.1371/journal.pone.0272269

Akila Devi, T.R., Javubar Sathick, K., Abdul Azeez Khan, A., & Arun Raj, L. (2023). Novel framework for improving the correctness of reference answers to enhance results of ASAG systems. *SN Computer Science*, *4*(4), Article 415. https://doi.org/10.1007/s42979-023-01682-8

Amur, Z.H., Hooi, Y.K., & Soomro, G.M. (2022). Automatic short answer grading (ASAG) using attention-based deep learning MODEL. In *2022 International Conference on Digital Transformation and Intelligence* (pp. 1-7). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICDI57181.2022.10007187

Badry, R.M., Ali, M., Rslan, E., & Kaseb, M.R. (2023). Automatic arabic grading system for short answer questions. *IEEE Access*, *11*, 39457-39465. https://doi.org/10.1109/ACCESS.2023.3267407

Benli, I., & İsmailova, R. (2018). Use of open-ended questions in measurement and evaluation methods in distance education. *International Technology and Education Journal*, *2*(1), 1-8.

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

Chan, S., Sathyamurthy, M., Inoue, C., Bax, M., Jones, J., & Oyekan, J. (2024). Integrating metadiscourse analysis with transformer-based models for enhancing construct representation and discourse competence assessment in l2 writing: A systemic multidisciplinary approach. *Journal of Measurement and Evaluation in Education and Psychology*, *15*(Special Issue), 318-347. https://doi.org/10.21031/epod.1531269

Chaudhari, R., & Patel, M. (2024). Deep learning in automatic short answer grading: A comprehensive review. *ITM Web of Conferences*, *65*, Article 03003. https://doi.org/10.1051/itmconf/20246503003

Chen, X., Zhou, Z., & Prado, M. (2025). ChatGPT-3.5 as an automatic scoring system and feedback provider in IELTS exams. *International Journal of Assessment Tools in Education*, *12*(1), 62-77. https://doi.org/10.21449/ijate.1496193

Chen, Y., Luo, J., Zhu, X., Wu, H., & Yuan, S. (2023). A cross-lingual hybrid neural network with interaction enhancement for grading short-answer texts. *IEEE Access, 11*, 37508-37514. https://doi.org/10.1109/ACCESS.2023.3260840

Çınar, A., İnce, E., Gezer, M., & Yılmaz, O. (2020). Machine learning algorithm for grading open-ended physics questions in Turkish. *Education and Information Technologies*, *25*(5), 3821-3844. https://doi.org/10.1007/s10639-020-10128-0

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dönmez, M. (2024). AI-based feedback tools in education: a comprehensive bibliometric analysis study. *International Journal of Assessment Tools in Education*, *11*(4), 622-646. https://doi.org/10.21449/ijate.1467476

Filighera, A., Ochs, S., Steuer, T., & Tregel, T. (2023). Cheating automatic short answer grading with the adversarial usage of adjectives and adverbs. *International Journal of Artificial Intelligence in Education, 34*, 616-646. https://doi.org/10.1007/s40593-023-00361-2

Garg, J., Papreja, J., Apurva, K., & Jain, G. (2022). Domain-specific hybrid BERT based system for automatic short answer grading. In *Proceedings of 2nd International Conference on Intelligent Technologies* (pp. 1-6). The Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/CONIT55038.2022.9847754

Ghavidel, H.A., Zouaq, A., & Desmarais, M.C. (2020). Using BERT and XLNET for the automatic short answer grading task. In H.C. Lane, S. Zvacek, & J. Uhomoibhi (Eds.), *Proceedings of the 12th International Conference on Computer Supported Education - (Volume 1)* (pp. 58-67). SciTePress. https://doi.org/10.5220/0009422400580067

Gomaa, W.H., Nagib, A.E., Saeed, M.M., Algarni, A., & Nabil, E. (2023). Empowering short answer grading: integrating transformer-based embeddings and BI-LSTM network. *Big Data and Cognitive Computing*, *7*(3), Article 122. https://doi.org/10.3390/bdcc7030122

Haller, S., Aldea, A., Seifert, C., & Strisciuglio, N. (2022). *Survey on automatic short answer grading with deep learning: From word embeddings to transformers*. arXiv. https://doi.org/10.48550/arXiv.2204.03503

Hasanah, U., Permanasari, A.E., Kusumawardani, S.S., & Pribadi, F.S. (2016, August). A review of an information extraction technique approach for automatic short answer grading. In *Proceedings of 2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering* (pp. 192-196). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICITISEE.2016.7803072

Jadidinejad, A.H., & Mahmoudi, F. (2014). Unsupervised short answer grading using spreading activation over an associative network of concepts / la notation sans surveillance des réponses courtes en utilisant la diffusion d'activation dans un réseau associatif de concepts. *Canadian Journal of Information and Library Science*, *38*(4), 287-303.

Katsaris, I., & Vidakis, N. (2021). Adaptive e-learning systems through learning styles: a review of the literature. *Advances in Mobile Learning Educational Research*, *1*(2), 124-145. https://doi.org/10.25082/AMLER.2021.02.007

Kurbanoğlu, N.I., & Olcaytürk, M. (2023). Investigation of the exam question types attitude scale for secondary school students: development, validity, and reliability. *Sakarya University Journal of Education*, *13*(2), 191-206. https://doi.org/10.19126/suje.1187470

Leacock, C., & Chodorow, M. (2003). C-rater: Automatic scoring of short-answer questions. *Computers and the Humanities*, *37*, 389-405. https://doi.org/10.1023/A:1025779619903

Li, X., Li, X., Chen, S., Ma, S., & Xie, F. (2022). Neural-based automatic scoring model for Chinese-English interpretation with a multi-indicator assessment. *Connection Science*, *34*(1), 1638-1653. https://doi.org/10.1080/09540091.2022.2078279

Lubis, F.F., Putri, A., Waskita, D., Sulistyaningtyas, T., Arman, A.A., & Rosmansyah, Y. (2021). Automatic short-answer grading using semantic similarity based on word embedding. *International Journal of Technology*, *12*(3), 571-581. https://doi.org/10.14716/ijtech.v12i3.4651

Mardini, G.I.D., Quintero, M.C.G., Viloria, N.C.A., Percybrooks, B.W.S., Robles, N.H.S., & Villalba, R.K. (2024). A deep-learning-based grading system (ASAG) for reading comprehension assessment by using aphorisms as open-answer-questions. *Education and Information Technologies*, *29*(4), 4565-4590. https://doi.org/10.1007/s10639-023-11890-7

Meurers, D., Ziai, R., Ott, N., & Kopp, J. (2011). Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In S. Padó, & S. Thater (Eds.), *Proceedings of the TextInfer 2011 workshop on textual entailment* (pp. 1–9). Association for Computational Linguistics. https://aclanthology.org/W11-2401/

Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In A. Lascarides, C. Gardent, & J. Nivre (Eds.), *Proceedings of the 12th Conference of the European chapter of the association for computational linguistics* (pp. 567-575). Association for Computational Linguistics. https://aclanthology.org/E09-1065.pdf

Nael, O., ElManyalawy, Y., & Sharaf, N. (2022). AraScore: a deep learning-based system for Arabic short answer scoring. *Array*, *13*, Article 100109. https://doi.org/10.1016/j.array.2021.100109

Nath, S., Parsaeifard, B., & Werlen, E. (2023, August 22-26). *Automatic short answer grading using BERT on German datasets* [Paper presentation]. 20th Biennial EARLI Conference, Thessaloniki, Greece.

Noyes, K., McKay, R.L., Neumann, M., Haudek, K.C., & Cooper, M.M. (2020). Developing computer resources to automate analysis of students' explanations of London dispersion forces. *Journal of Chemical Education*, *97*(11), 3923-3936. https://doi.org/10.1021/acs.jchemed.0c00445

Padó, U. (2016). Get semantic with me! the usefulness of different feature types for short-answer grading. In Y. Matsumoto, & R. Prasad (Eds.), *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical Papers* (pp. 2186-2195). The COLING 2016 Organizing Committee. https://aclanthology.org/C16-1206/

Page, E.B. (1967). Grading essays by computer: Progress report. *Proceedings of the Invitational Conference on Testing Problems*, 87-100.

Ramineni, C., & Williamson, D.M. (2013). Automatic essay scoring: psychometric guidelines and practices. *Assessing Writing*, *18*(1), 25-39. https://doi.org/10.1016/j.asw.2012.10.004

Riyanto, S., Imas, S.S., Djatna, T., & Atikah, T.D. (2023). Comparative analysis using various performance metrics in imbalanced data for multi-class text classification. *International Journal of Advanced Computer Science and Applications*, *14*(6). https://doi.org/10.14569/IJACSA.2023.01406116

Salim, H.R., De, C., Pratamaputra, N.D., & Suhartono, D. (2022). Indonesian automatic short answer grading system. *Bulletin of Electrical Engineering and Informatics*, *11*(3), 1586-1603. https://doi.org/10.11591/eei.v11i3.3531

Saunders, D.R., Bex, P.J., Rose, D.J., & Woods, R.L. (2014). Measuring information acquisition from sensory input using automatic scoring of natural-language descriptions. *PLoS ONE*, *9*(4), Article e93251. https://doi.org/10.1371/journal.pone.0093251

Sawatzki, J., Schlippe, T., & Benner-Wickner, M. (2021). Deep learning techniques for automatic short answer grading: predicting scores for English and German answers. In E. Cheng, R.B. Koul, T. Wang, & X. Yu (Eds.), *Proceedings of 2021 2nd international conference on artificial intelligence in education technology* (pp. 65-75). Springer. https://doi.org/10.1007/978-981-16-7527-0_5

Sayeed, M.A., & Gupta, D. (2022). Automate descriptive answer grading using reference based models. In *Proceedings of 2022 OITS international conference on information technology* (pp. 262-267). The Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/OCIT56763.2022.00057

Schleifer, A.G., Klebanov, B.B., Ariely, M., & Alexandron, G. (2023). Transformer-based Hebrew NLP models for short answer scoring in biology. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications* (pp. 550-555). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.46

Seker, A., Bandel, E., Bareket, D., Brusilovsky, I., Greenfeld, R., & Tsarfaty, R. (2022). AlephBERT: Language model pre-training and evaluation from sub-word to sentence level. In *Proceedings of the 60th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 46-56). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.4

Siddiqi, R., Harrison, C.J., & Siddiqi, R. (2010). Improving teaching and learning through automatic short-answer marking. *IEEE Transactions on Learning Technologies*, *3*(3), 237-249. https://doi.org/10.1109/TLT.2010.4

Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V., & Arora, R. (2019, November). Pre-training BERT on domain resources for short answer grading. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 6071-6075). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1628

Şenel, S., & Şenel, H.C. (2021). Remote assessment in higher education during COVID-19 pandemic. *International Journal of Assessment Tools in Education*, *8*(2), 181-199. https://doi.org/10.21449/ijate.820140

Tulu, C.N., Özkaya, O., & Orhan, U. (2021). Automatic short answer grading with semspace sense vectors and malstm. *IEEE Access, 9*, 19270-19280. https://doi.org/10.1109/ACCESS.2021.3054346

Uto, M., & Uchida, Y. (2020). Automatic short-answer grading using deep neural networks and item response theory. In I.I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millan (Eds.), *Proceedings of the 21th International Conference on Artificial Intelligence in Education* (pp. 334-339). Springer International Publishing. https://doi.org/10.1007/978-3-030-52240-7_61

Uyar, A.C., & Büyükahıska, D. (2025). Artificial intelligence as an automated essay scoring tool: a focus on ChatGPT. *International Journal of Assessment Tools in Education*, *12*(1), 20-32. https://doi.org/10.21449/ijate.1517994

Uysal, I., & Doğan, N. (2021). How reliable is it to automatically score open-ended items? An application in the Turkish language. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(1), 28-53. https://doi.org/10.1007/978-3-030-52240-7_61

Westera, W., Dascalu, M., Kurvers, H., Ruseti, S., & Trausan-Matu, S. (2018). Automatic essay scoring in applied games: Reducing the teacher bandwidth problem in online training. *Computers & Education*, *123*, 212-224. https://doi.org/10.1016/j.compedu.2018.05.010

Xie, M.K., Xiao, J., Liu, H.Z., Niu, G., Sugiyama, M., & Huang, S.J. (2023). Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning. *Advances in Neural*

Information Processing Systems, *36*, 25731-25747. https://doi.org/10.48550/arXiv.2305.02795

Yang, X., Huang, J.Y., Zhou, W., & Chen, M. (2022). *Parameter-efficient tuning with special token adaptation*. arXiv. https://doi.org/10.48550/arXiv.2210.04382

Yıldırım, O., & Demir, S.B. (2022). Inside the black box: Do teachers practice assessment as learning?. *International Journal of Assessment Tools in Education*, *9*(Special Issue), 46-71. https://doi.org/10.21449/ijate.1132923

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement*, *76*(2), 280-303. https://doi.org/10.1177/0013164415590022

Zesch, T., Horbach, A., & Zehner, F. (2023). To score or not to score: Factors influencing performance and feasibility of automatic content scoring of text responses. *Educational Measurement: Issues and Practice*, *42*(1), 44-58. https://doi.org/10.1111/emip.12544

Zhang, L., & Copus, B. (2023). A Study of Compressed Language Models in Social Media Domain. *The International FLAIRS Conference Proceedings*, *36*(1). https://doi.org/10.32473/flairs.36.133056

Zhang, M., Johnson, M., & Ruan, C. (2024). Investigating sampling impacts on an LLM-based AI scoring approach: Prediction accuracy and fairness. *Journal of Measurement and Evaluation in Education and Psychology*, *15*(Special Issue), 348-360. https://doi.org/10.21031/epod.1561580

Zhu, X., Wu, H., & Zhang, L. (2022). Automatic short-answer grading via BERT-based deep neural networks. *IEEE Transactions on Learning Technologies*, *15*(3), 364-375. https://doi.org/10.1109/tlt.2022.3175537

Zimmerman, W.A., Kang, H.B., Kim, K., Gao, M., Johnson, G., Clariana, R., & Zhang, F. (2018). Computer-automatic approach for scoring short essays in an introductory statistics course. *Journal of Statistics Education*, *26*(1), 40-47. https://doi.org/10.1080/10691898.2018.1443047

APPENDIX

Appendix 1. *Physics dataset questions and scoring keys (Çınar et al., 2020).*

Question ID	Question Text	Scoring Key
Q1	Sinan is preparing the report of the experiment they did in the physics laboratory class that day. The question in the test report is: “If there is a single bulb in an electrical circuit, in which case two bulbs of the same power are connected in series, compare the brightness of the bulbs and explain the reason for your answer.” Sinan rethinks their experiments in the laboratory and answers the question as follows: “While there is only one light bulb in an electrical circuit, the light bulb is quite bright. When a second bulb is added serially to this bulb, both bulbs light up equally and less brightly. The reason for this is that the single bulb uses all the current flowing through the circuit. When there are two bulbs in the circuit, the current is shared by the bulbs and therefore they are equal but less bright.” Sinan thinks he understands very well what is happening in the experiment, but he makes a big mistake. Can you find this error?	0 point: If student did not answer or gave an incorrect answer. 1 point: Sinan thinks that the current passing through both bulbs and the bulbs share this current. However, the same current passes first through the first bulb and then through the other. The two bulbs are dimmer than a single bulb. Because the increased resistance in the circuit reduces the current. Sinan; considers that the bulbs are connected in parallel.
Q2	You drop an apple from a certain height without speed and it hits the ground with a certain velocity. Assume that you cut the same apple in half and drop half of it at the same height. (Assume no air friction, please express your answers only verbally without writing formulas) a) What is the speed of the cut apple in half according to the speed of the whole apple? b) Explain the reason by considering the factors that affect its speed for the answer to option a. c) Explain the reason taking into account the energy conservation laws for your answer to option a.	a) 0 point: If student did not answer or gave incorrect answer (in this case the answers of b and c are ignored). 1 point: Speed does not change. / Speeds are the same. b) 0 point: If student did not answer or gave incorrect answer. 2 points (if option a is correct): only if height is specified in option b, 2 points (if option a is correct): only if mass is specified in option b, 3 points (if option a is correct): In option b, if both height and mass are specified: The objects that are allowed to fall free from the same height have different masses, but their velocities are the same. Because the velocity of an object that makes free fall movement is related to height but it is independent of mass. c) 0 point: If student did not answer or gave incorrect answer. 4 points (if option a and b are correct): According to energy conservation laws, if an object has only potential energy when it is at a certain height and the object is allowed to fall free from the height, the height decreases as it falls to the ground as soon as it is allowed to fall but the object starts to accelerate. In this case, the kinetic energy increases while the potential energy decreases. As soon as the object hits the ground, it has only kinetic energy increased and the total mechanical energy is preserved.

Question ID	Question Text	Scoring Key
Q3	What is Mechanical Energy? Please explain. (Please express your answers only verbally without writing formulas)	0 point: If student did not answer or gave incorrect answer. 1 point: Mechanical energy is the sum of potential and kinetic energy for conservative systems.
Q4	What does scientific Work mean? Please explain. (Please express your answers only verbally without writing formulas)	0 point: If student did not answer or gave incorrect answer. 1 point: Work is applying force to an object in a certain direction and moving that object in the direction of the applied force.