MOJES MALAYSIAN ONLINE JOURNAL OF EDUCATIONAL SCIENCES APRIL 2026, 14 (2)

INVESTIGATION OF ARTIFICIAL INTELLIGENCE SUPPORTED AUTOMATIC ASSESSMENT AND EVALUATION SYSTEMS IN SCIENCE EDUCATION IN TERMS OF QUALITY

***Gül İrem Özen[1]**
Mahmut Selvi[2]
[1] Independent Researcher, Türkiye
[2] Department of Science Education, Faculty of Education, Gazi University, Ankara, Türkiye
**guliremozen@gmail.com*

ABSTRACT

This study examines the quality dimensions of artificial intelligence-based automated assessment systems in science education in terms of validity, reliability, and conceptual understanding. A DeBERTa-v3-based multiple-choice question answering model was trained using the SciQ dataset. The model was evaluated through accuracy analysis, justification-based semantic similarity, and comparative performance analysis. The findings indicate that the model achieved high accuracy (90.4%) and demonstrated strong alignment between selected answers and supporting texts. Beyond accuracy, the system shows potential to assess students' conceptual understanding through justification-based analysis. The results suggest that AI-supported systems can enhance the quality of both formative and summative assessment practices by providing meaningful, explanation-based insights into student reasoning. However, considerations related to explainability, pedagogical alignment, and ethical use remain critical for effective classroom implementation. These findings contribute to bridging the gap between technical model performance and pedagogical assessment practices in science education. This study provides a novel contribution by integrating justification-based semantic analysis into AI-supported assessment, offering a multidimensional evaluation framework for science education.

Keywords: *Artificial intelligence, automated assessment, validity and reliability, assessment quality, DeBERTa-v3*

INTRODUCTION

In the evaluation of student achievement in science education, it is not enough to give the correct answer; it is of great importance that students can justify their answers on scientific grounds (Messick, 1995). This approach allows students to go beyond the acquisition of superficial knowledge and develop in-depth comprehension and scientific reasoning skills (Brickhouse et al., 2000; Doshi-Velez & Kim, 2017). While the justification process supports students' scientific argumentation, critical thinking, and decision-making skills, it also plays an important role in quality assurance in assessment and evaluation practices (Ganesan et al., 2023; Shute, 2008). For this reason, to evaluate learning outcomes in science education validly and comprehensively, it is imperative to take into account the justification provided by the student as well as the correct answer (Shute, 2008).

Artificial intelligence (AI) technologies have the potential to transform teaching and learning processes in education (Holmes et al., 2019; Luckin et al., 2016). In particular, automated assessment systems can analyze student performance quickly, objectively, and consistently, thus reducing teachers' assessment burden (Baker & Inventado, 2014; Heffernan & Heffernan, 2014). However, the potential of these systems to analyze not only the ability to select the correct answer but also the scientific

justifications supporting the student's answer is increasing (Reimers & Gurevych, 2019; Wang & Degol, 2017). Thus, student responses can be evaluated not only for technical accuracy but also for pedagogical relevance and explainability (Doshi-Velez & Kim, 2017; Farrow, 2023).

With the widespread use of automated assessment systems, AI-supported systems offer important opportunities, especially in areas that require complex and conceptual justification, such as science education (Bull & Kay, 2016; Holmes et al., 2019). In addition to providing high accuracy rates in the evaluation of multiple-choice or short-answer questions, these systems have the potential to examine the scientific accuracy and consistency of student responses (He & Yih, 2023; Reimers & Gurevych, 2019). Thus, learning outcomes can be analyzed more holistically in both summative and formative assessment practices (Kizilcec, 2022; Williamson & Eynon, 2020).

In this framework, the question of to what extent AI-based systems can provide valid, reliable, and explainable assessment in the context of science education is becoming increasingly critical (Baker & Siemens, 2014; Kizilcec, 2022; Williamson & Eynon, 2020). This is because not only technical accuracy but also pedagogical meaningfulness, context sensitivity, and qualities that support student learning are important components that determine the quality of assessment tools (Hattie & Timperley, 2007; Lai & Viering, 2012; Pellegrino et al., 2001). Therefore, a comprehensive examination of AI-based assessment and evaluation systems at both technical and pedagogical levels provides important clues on how and under what conditions these tools can be used effectively in education (Bull & Kay, 2016; Holmes et al., 2019; Kay et al., 2022).

In this context, this study aims to examine the quality dimensions of AI-supported automated assessment-evaluation systems in science education within the framework of validity, reliability, and explainability. Accordingly, the capacity of a model developed based on the SciQ dataset to determine the correct answer (validity), its ability to establish a meaningful relationship with the support text (reliability), and its usability in teaching environments (measurement quality) will be evaluated. Nowadays, there is an increasing need for systems that support individualized learning, analyze student thinking styles, and provide diagnostic data to the teacher (Ganesan et al., 2023; Holmes et al., 2019). This study aims to contribute to the development of ethical, explicable, and pedagogically meaningful assessment and evaluation systems in science education.

In this context, the system is designed not only to identify correct answers but also to support teachers in evaluating students' reasoning processes through justification-based semantic analysis. This enables teachers to interpret student responses beyond correctness and to generate more targeted and meaningful feedback.

This study contributes to the literature in three significant ways. First, it integrates transformer-based AI models with core assessment principles such as validity, reliability, and explainability. Second, it introduces a justification-based evaluation approach that goes beyond accuracy by capturing students' conceptual understanding. Third, it provides empirical evidence on the potential of AI-supported systems to enhance the quality of assessment practices in science education. Unlike previous studies that focus primarily on accuracy, this study provides a multidimensional evaluation framework that integrates validity, reliability, and explainability within a pedagogical context.

The research questions were formulated in alignment with the operational definitions of validity, reliability, and conceptual understanding.

1. RQ1. To what extent does the AI-based automated evaluation system developed using the SciQ dataset demonstrate validity, as measured by its accuracy in identifying correct answers to science questions?
2. RQ2. To what extent does the system demonstrate reliability in evaluating justification-based responses, as measured by the degree of alignment between the most semantically similar option and the correct answer?

3. RQ3. To what extent does the system capture conceptual understanding through justification-based semantic alignment within the SciQ dataset, and how does it contribute to the quality of formative and summative assessment practices?

Theoretical Framework

Measurement, Evaluation and Quality. Assessment and evaluation in science education is a fundamental process to accurately and comprehensively determine students' scientific knowledge, skills, and attitudes (Pellegrino et al., 2001; Popham, 2017; Shavelson, 2006). Effective assessment and evaluation practices should comprehensively assess not only correct or incorrect answers, but also students' thinking processes, scientific reasoning, and conceptual understanding (Brookhart, 2013; Osborne, 2014; Shepard, 2000). Such approaches allow students to demonstrate their critical thinking and scientific argument development competencies and provide diagnostic feedback to teachers (Sadler, 2004; Wilson & Bertenthal, 2005).

The validity and reliability levels of measurement tools play a critical role in determining not only individual achievement but also the quality of instructional design and learning environments (Kane, 2013; Messick, 1995; Nitko & Brookhart, 2014). Especially in science education, students frequently fall into conceptual misconceptions, which makes it compulsory to measure abstract and multidimensional thinking skills (Novak, 2010; Treagust, 1988). This situation reveals that classical true-false-based measurement approaches are often insufficient and the necessity of more explanatory and meaning-centered measurement systems (Black & Wiliam, 2009; Ruiz-Primo & Furtak, 2006).

The Place of AI in Measurement. AI is being applied more and more widely in the field of educational technologies and is creating radical transformations in teaching, learning, and assessment and evaluation processes (Chen et al., 2020; Holmes et al., 2019; Luckin et al., 2016). AI offers the opportunity to provide personalized feedback by analyzing student performance in real time and dynamically adapting instructional materials according to individual needs (Kay et al., 2022; Kizilcec, 2022; Zawacki-Richter et al., 2019). In particular, natural language processing and machine learning-based systems are gradually developing in terms of evaluating students' written responses, analyzing their reasoning structures and modelling their cognitive processes (Baker & Inventado, 2014; Ganesan et al., 2023; Heffernan & Heffernan, 2014).

Automated assessment systems reduce the burden on teachers and increase the efficiency of assessment and evaluation processes by providing fast, consistent, and objective evaluation of student responses (Shute & Rahimi, 2021; VanLehn, 2011; Williamson & Eynon, 2020). In disciplines such as science that require hypothesizing, establishing cause-effect relationships, and multidimensional justification, AI-supported assessment systems offer significant potential not only in terms of technical efficiency but also pedagogical precision, conceptual validity, and explainability (Bull & Kay, 2016; Doshi-Velez & Kim, 2017; Farrow, 2023).

Automated Assessment Systems and Processes. Automated assessment systems are AI-based tools developed to analyze student responses quickly, consistently, and objectively (Shute, 2008; Zawacki-Richter et al., 2019). The basic components of these systems include data collection, preprocessing, model training, evaluation, and instant feedback stages (Chen et al., 2020; Heffernan & Heffernan, 2014; Kay et al., 2022). In the first stage of the process, student responses are collected in digital environments and transformed into meaningful text structures using natural language processing (NLP) techniques (Jurafsky & Martin, 2020; Manning et al., 2014).

Then, algorithms trained using large-scale language models or deep learning-based approaches operate on classification or ranking tasks to distinguish between correct and reasoned responses (Devlin et al., 2019; Khashabi et al., 2020; Raffel et al., 2020). These models focus on analyzing not only the accuracy of the result, but also the scientific reasoning in the student's response (Farrow, 2023; Ganesan et al., 2023). In the final stage of the process, the system supports learning processes by providing

personalized and instant feedback on student performance (Baker & Inventado, 2014; Holmes et al., 2019).

The use of these technologies in the context of science education not only accelerates the assessment process but also provides the opportunity to analyze students' conceptual errors, thinking patterns, and scientific reasoning strategies in more depth (Bull & Kay, 2016; Kizilcec, 2022; Luckin et al., 2016).

Validity and Reliability Approaches. Quality in measurement and assessment processes is generally associated with basic criteria such as validity, reliability, and explainability (Kane, 2013; Nitko & Brookhart, 2014; Pellegrino et al., 2001). Validity refers to the extent to which a measurement tool accurately measures the targeted knowledge, skill, or competence for a specific purpose (Messick, 1995; Shepard, 2000). The validity of an assessment tool requires that it produces results that are meaningful, interpretable, and suitable for decision-making about the concept it measures (American Educational Research Association [AERA], 2014; Cohen et al., 2018).

Reliability is defined as the degree to which a measurement tool gives similar results under different conditions; in this context, it is related to the concepts of consistency and repeatability (Bachman & Palmer, 2010; Crocker & Algina, 1986; Shavelson, 2006). When these two criteria are evaluated together, they constitute the main indicators of a measurement tool's potential to ensure quality in education (Brookhart, 2013; Wang & Degol, 2017).

In particular, the integration of AI-supported automated assessment systems into science education requires the revision of validity and reliability criteria (Ganesan et al., 2023; Zawacki-Richter et al., 2019). In such systems, not only selecting the correct answer but also analyzing student responses based on justification plays a critical role in the construction of a valid measurement structure (Bull & Kay, 2016). Therefore, systematic monitoring of both technical accuracy and pedagogical meaningfulness in AI-based systems is of great importance in terms of obtaining quality assessment and evaluation outputs in education (Kay et al., 2022; Williamson & Eynon, 2020).

Quality Dimension in Science Education and AI Applications. For AI-based assessment systems to be used effectively and reliably in education, the quality criteria of these systems should be carefully examined (Birenbaum & Tatsuoka, 2013; Kay et al., 2022; Williamson & Eynon, 2020). As with all types of measurement tools used in education, validity and reliability are among the basic quality criteria in these systems. While validity means that the system measures the targeted knowledge or skill accurately and meaningfully, reliability ensures that it can provide consistent results when repeated under the same conditions (Kane, 2013; Messick, 1995; Nitko & Brookhart, 2014).

AI systems have the capacity to analyze not only the accuracy of student responses but also the justification processes underlying these responses, especially thanks to their natural language processing and machine learning-based structures (Ganesan et al., 2023; Shute & Rahimi, 2021). In this context, validity in AI-supported assessment systems is not limited to determining the correct answer; students' scientific thinking, argumentation competences, and conceptual explanations are also part of this validity structure (Bull & Kay, 2016; Sotomayor et al., 2021).

At this point, science education-focused multiple-choice question sets, such as the SciQ dataset, provide an important platform for testing both the knowledge level and reasoning analysis ability of AI systems (Khashabi et al., 2018; Khot et al., 2020). In addition, factors such as the scope of the data sets in which these systems are developed, the nature of the questions, and the structure of the supporting texts also have a decisive effect on the reliability of the measurement results (Zawacki-Richter et al., 2019). The diversity of the data set directly affects the ability of the AI model to recognize different thinking patterns and increases the explainability and adaptability of the measurement system (Doshi-Velez & Kim, 2017; Kizilcec, 2022).

As a result, addressing quality criteria with a holistic approach increases the pedagogical value of AI-based assessment systems in science education and enables the development of reliable digital

environments that support teacher and student interactions (Farrow, 2023; Holmes et al., 2019). In this context, this study aims to examine how an AI-based automated assessment system developed based on the SciQ dataset offers a quality profile in terms of validity and reliability.

LITERATURE REVIEW

Automatic Evaluation Systems with AI

In recent years, AI-based automatic assessment systems have become increasingly widespread in the field of education with the aim of accelerating, standardizing, and making assessment processes more objective (Kizilcec, 2022; Luckin et al., 2016; Zawacki-Richter et al., 2019). These systems can automatically score student responses by analyzing open-ended and multiple-choice questions (Baker & Inventado, 2014; Ganesan et al., 2023; VanLehn, 2011). This reduces the assessment burden on teachers while increasing transparency and efficiency in the measurement process (Heffernan & Heffernan, 2014; Shute & Rahimi, 2021).

One of the most important advantages of automated systems is that they save time and strengthen the concept of 'assessment quality' by reducing human subjectivity and inconsistency (Bull & Kay, 2016; Shute, 2008; Williamson & Eynon, 2020). Especially in justification-based disciplines such as science education, advanced models not only find the correct answer but also analyze the justification structures in student responses (Sotomayor et al., 2021). This capability enhances the explainability of assessment systems and enables teachers to gain deeper insights into students' thinking processes (Doshi-Velez & Kim, 2017; Farrow, 2023; Holmes et al., 2019).

Assessment and Quality in Science Education

The fundamental purpose of assessment in science education is not limited to measuring students' knowledge levels; it also aims to evaluate higher-order thinking skills such as scientific reasoning, argumentation, and conceptual understanding (Lehrer & Schauble, 2006; Osborne et al., 2004). In this context, effective assessment practices should focus on understanding how students think throughout the learning process and how they justify their thought processes (Furtak et al., 2016).

It is not enough for assessment to merely provide test results; an effective measurement and assessment process must fulfil both formative and summative functions (Black & Wiliam, 2009; Heritage, 2010). Formative assessment supports student development by providing continuous feedback on the learning process, while summative assessments are used to document academic achievement and make systemic decisions (Pellegrino et al., 2001; Shepard, 2000).

A quality assessment process depends not only on determining the 'correct' answer, but also on the existence of systems that can explain why students gave that answer (Brickhouse et al., 2000; Doshi-Velez & Kim, 2017). This level of explainability enhances the pedagogical value of assessment practices and facilitates teachers' insight into students' thinking processes (Holmes et al., 2019; Shute & Rahimi, 2021).

In this context, there is a growing need for research that reveals the extent to which AI-based assessment systems align with quality criteria specific to science education (Bull & Kay, 2016; Ganesan et al., 2023; Zawacki-Richter et al., 2019). For these systems to be effectively integrated into educational environments, analyses based not only on accuracy rates but also on multidimensional quality criteria such as validity, reliability, and explainability are required (Kane, 2013; Williamson & Eynon, 2020).

Concepts of Validity and Reliability

Validity refers to the degree to which a measurement tool accurately and meaningfully measures the targeted knowledge, skill, or trait, while reliability refers to the consistency and repeatability of measurement results (Kane, 2013; Messick, 1995; Shavelson, 2006). In the context of educational technologies, particularly in AI-supported assessment systems, these two fundamental criteria are central to the concept of 'assessment quality' and are considered indispensable for the reliable use of such systems (AERA, 2014; Nitko & Brookhart, 2014; Pellegrino et al., 2001).

AI systems, particularly when used with models that adhere to the principle of explainability, not only determine the accuracy of student responses but also, can analyze the conceptual depth, scientific justification level, and thought patterns (Doshi-Velez & Kim, 2017; Holmes et al., 2019; Wang & Degol, 2017). Thus, not only numerical success, but also students' thought processes can be included in the measurement and evaluation scope (Bull & Kay, 2016; Shute & Rahimi, 2021).

However, it should be noted that validity and reliability are not only related to model performance but also to numerous factors such as the quality of the data set used, model architecture, comprehensiveness of training data, and pedagogical compatibility of evaluation criteria (Cohen et al., 2018; Ganesan et al., 2023; Kizilcec, 2022). At this point, the discussion of the model's usability not only in terms of technical accuracy but also in terms of explainability in a pedagogical context and its contribution to the learning process is gaining increasing importance in the literature (Williamson & Eynon, 2020; Zawacki-Richter et al., 2019).

SciQ Dataset and MCQA Studies

The SciQ dataset (Welbl et al., 2017) consists of multiple-choice questions suitable for middle school-level science, accompanied by supporting texts. Each question has one correct answer and three carefully crafted distractors. This structure allows the dataset to measure not only knowledge recall but also cognitive processes such as establishing relationships between pieces of information, reasoning, and context analysis (Khashabi et al., 2018; Sugawara et al., 2018). SciQ has thus become an important benchmark dataset for multiple-choice question answering (MCQA) problems (Liu et al., 2019; Rajpurkar et al., 2016).

Effective results have been obtained on SciQ, especially with models based on transformer architecture. BERT (Devlin et al., 2019) and its version adapted to scientific literature, SciBERT (Beltagy et al., 2019), stand out for their ability to analyze contextual relationships between texts. The fact that SciBERT has been trained on scientific journals and academic texts makes it more meaningful for questions in the context of science education (Arora et al., 2020; Lo et al., 2023). This feature is effective in ensuring not only superficial accuracy but also conceptual validity (Pavlick & Kwiatkowski, 2019).

In addition, SciQ offers a structure that is suitable not only for summative assessment (e.g., achievement tests) but also for formative analyses (e.g., identifying types of student misunderstandings, analyzing reasoning structures) (Clark et al., 2018; Ganesan et al., 2023). From this perspective, the integration of AI-based MCQA systems developed based on data sets such as SciQ into science education offers a multi-layered evaluation framework that addresses dimensions such as technical accuracy, pedagogical explainability, and measurement quality (Bull & Kay, 2016; Shute & Rahimi, 2021; Zawacki-Richter et al., 2019).

Evaluation of MCQA Models in Terms of Validity and Reliability

Validity refers to the degree to which an assessment tool accurately measures the construct it is intended to measure (Kane, 2013; Messick, 1995). Especially in educational technologies and MCQA (multiple-choice question answering) models, validity is not limited to selecting the correct answer, but is also related to how accurately the model interprets scientific content within context (AERA, 2014; Cohen et al., 2018). For example, in models such as UnifiedQA, not including SciQ test data in the training set strengthens the validity of the assessment, thereby enabling the system's generalization ability to be measured more reliably (Khashabi et al., 2020).

Indeed, the fact that the UnifiedQA model achieved 63% accuracy without seeing SciQ data and then reached 78% accuracy after training demonstrates the development of structural validity and context sensitivity (Khashabi et al., 2020; Sugawara et al., 2018). This type of context awareness enables analyses that are closer to students' ways of thinking (Bull & Kay, 2016; Shute & Rahimi, 2021).

The data set itself must also be carefully examined for validity. The SciQ dataset covers basic disciplines such as physics, chemistry, and biology and contains questions based on middle school-level curricula (Welbl et al., 2017). Therefore, its content validity is strong (Clark et al., 2018). Additionally, the

supporting paragraph provided for each question ensures that students use evidence-based reasoning rather than rote knowledge, thereby supporting the construct validity dimension (Arora et al., 2020; Pavlick & Kwiatkowski, 2019).

In some models (e.g., SBERT-based systems), the accuracy of the response can be checked not only by the label but also by the semantic proximity (cosine similarity) with the supporting paragraph (Ganesan et al., 2023; Reimers & Gurevych, 2019). This enables automatic validity control in machine-based evaluations (Doshi-Velez & Kim, 2017).

Reliability refers to a measurement tool's ability to produce repeatable and consistent results (Crocker & Algina, 1986; Shavelson, 2006). The fact that MCQA models produce similar results in different runs is particularly important in terms of educational reliability. In studies related to the UnifiedQA model, the accuracy rate on SciQ being reported with an error margin of ±2.8% across three different runs supports test-retest reliability (Khashabi et al., 2020).

However, large language models (LLMs) can exhibit some inconsistencies in prompt-based studies. For example, the GPT-3.5 model showed accuracy rates of 45% and 41% in two different runs of the same question, with response consistency calculated at 74% (Borji, 2023). This indicates that careful evaluation is required in an educational context due to the low level of internal consistency (Williamson & Eynon, 2020; Zawacki-Richter et al., 2019). To mitigate such issues, strategies such as ensemble learning, self-consistency, or multiple response generation and voting are proposed (Liang et al., 2022; Wang et al., 2022).

As a result, evaluating MCQA models in terms of validity and reliability should not be limited to looking at accuracy rates; how the model accesses information, how consistently it uses information, and how well it aligns with the content should also be analyzed (Holmes et al., 2019; Pellegrino et al., 2001). Such analyses are critical for determining the extent to which AI systems can be used in science education, both in terms of technical suitability and pedagogical applicability.

Evaluation in Terms of Explainability

The explainability of AI systems used in education is not only a technical requirement but also a pedagogical necessity (Doshi-Velez & Kim, 2017; Miller, 2019). The ability of an assessment model to transparently explain why it gave a particular answer enables students to participate confidently in the learning process and teachers to interpret the model output in a meaningful way (Holmes et al., 2019; Ribeiro et al., 2016). This requires that, especially in scientific questions, not only the 'correct' answer but also the thought process leading to that answer be visible (Chen et al., 2020; Suresh & Guttag, 2021).

The SciQ dataset (Welbl et al., 2017) provides contextual data support for AI systems' explanation generation processes by offering a supporting paragraph for each question. This structure enables the model to explain its responses not with arbitrary justifications but by linking them to the paragraph. For example, Ganesan et al. (2023) contributed to explainability by comparing student responses with supporting paragraphs in their BERT-based automatic answer verification system. This approach represents an evaluation approach that measures not only the result but also the process (Zhou et al., 2022).

The 'chain-of-thought' method, which has gained prominence in recent years, is used as a strategy to increase explainability (Wei et al., 2022). In this technique, the model not only produces the answer but also presents the logical steps it used to reach the answer in writing. For example, Allard et al. (2024) equipped the LLaMA-SciQ model with chain-of-thought capability using the AQuA-RAT dataset, enabling the model to produce both the process and the result. This structure allows students to learn not only 'what' but also 'how' (Rajani et al., 2019).

To increase explainability, the internal structures of the model, namely attention weights, can be visualized (Clark et al., 2019). However, these techniques may not always align with the pedagogical

explanation of model decisions (Belinkov & Glass, 2019). Therefore, rationale generation methods based on directly generating justifications within the output both improve model performance and provide users with more meaningful information (Narang et al., 2020; Wang et al., 2022).

As a result, explainable AI systems not only automate the assessment process but can also play an instructive role in learning processes (Popenici & Kerr, 2017; Xie et al., 2021). While students can develop their own thinking strategies by examining the model's solution steps, teachers can intervene by evaluating the system's rationale (Luckin et al., 2016). Therefore, explainability is not only a measure of accuracy but also a determining indicator of usability in education (Williamson & Eynon, 2020).

Model Comparisons: Evaluation in Terms of Performance, Reliability, and Explainability

Studies conducted on the SciQ dataset reveal how various model types perform in terms of evaluation quality and educational applicability (Ganesan et al., 2023; Khashabi et al., 2020; Welbl et al., 2017). In this section, prominent MCQA models are compared in terms of accuracy, validity, reliability, and explainability.

Accuracy and Performance Comparison

The models' performance on SciQ varies greatly. For example:

- UnifiedQA-Large achieved an accuracy rate of 78.7% with a T5-based encoder-decoder structure (Khashabi et al., 2020).
- SBERT + Cosine Similarity provided 74.9% accuracy based on semantic similarity evaluation (Ganesan et al., 2023).
- Siamese BERT achieved 84.5% accuracy thanks to its dual structure that uses similarity learning (Ganesan et al., 2023).
- LLaMA 7B (fine-tuned) demonstrated the highest success with 90.4% accuracy as an LLM (Allard et al., 2024).

These differences vary depending on the capacity of the architectures used, the quality of the training data, and the model's fine-tuning strategies.

Evaluation in Terms of Validity

Validity questions whether the model actually measures the skill it is supposed to measure (Messick, 1995). The fact that UnifiedQA achieves 63.4% accuracy when used without any training on SciQ test data is a strong indicator of the model's generalization capacity (Khashabi et al., 2020). In contrast, although the LLaMA-SciQ model offers higher accuracy because it is fine-tuned specifically to the training data, validity must be carefully evaluated due to the risk of data leakage (Allard et al., 2024). In SBERT-based models, the model's determination of the answer by matching the text with the supporting paragraph directly supports content validity (Reimers & Gurevych, 2019).

Evaluation in Terms of Reliability

Reliability refers to the consistency of the model's outputs. In studies conducted for UnifiedQA, a 95% confidence interval was achieved with a deviation of ±2.8 points in tests conducted under three different conditions (Khashabi et al., 2020). This shows that the model is resistant to random deviations.

In contrast, models such as GPT-3.5 achieved 74% consistency between prompt-based outputs, but the rate of different responses reached 26% in two conditions (Thorp, 2023). This situation can reduce reliability, especially in classroom use.

Evaluation in Terms of Explainability

In educational technologies, it is important not only to give the correct answer but also to explain how that answer was reached. The SBERT + Cosine and Siamese BERT models ensure explainability by evaluating the semantic similarity between the supporting paragraph and the student's answer (Ganesan et al., 2023).

At a more advanced level, models such as LLaMA-SciQ offer significantly higher explainability because they first generate the solution process and then provide the answer using the chain-of-thought technique (Allard et al., 2024; Wei et al., 2022). This allows students to focus not only on the result but also on the process.

Table 1
Comparative Performance of AI Models on the SciQ Dataset

Model name	Architecture	Accuracy (%)	Validity	Reliability	Explainability
UnifiedQA-Large	T5-based	78.7	High	Moderate–High	Low–Moderate
SBERT + Cosine sim.	BERT-based	74.9	Moderate–High	High	Moderate–High
Siamese BERT	Siamese BERT	84.5	High	High	Moderate
LLaMA-SciQ (7B FT)	LLM	90.4	Moderate	Moderate	High

Note. Accuracy values are based on the test set of the SciQ dataset. FT = fine-tuned. Data adapted from Khashabi et al. (2020), Ganesan et al. (2023), and Allard et al. (2024).

As presented in Table 1, different AI models demonstrate varying trade-offs between accuracy, reliability, and explainability, highlighting the multidimensional nature of assessment quality.

The DeBERTa-v3-based model used in this study differs from previous AI-based assessment systems in that it potentially justify students' answers as well as provide the correct answer (Doshi-Velez & Kim, 2017; Ganesan et al., 2023). The model aims to fill gaps in the literature by being evaluated based on additional criteria such as pedagogical meaningfulness and explainability, in addition to validity and reliability criteria (Khashabi et al., 2020; Shute & Rahimi, 2021). This study goes beyond the limitations of previously widely used models, offering robust context analysis and justification-based accuracy analysis in the context of science education (Farrow, 2023; Holmes et al., 2019).

METHODOLOGY

Data Set

In this study, the SciQ dataset was used to evaluate multiple-choice question answering (MCQA) models in the context of science education. SciQ is an open-access dataset developed by the Allen Institute for AI, containing a total of 13,679 four-choice multiple-choice science questions (Welbl et al., 2017). The questions cover fundamental science fields such as physics, chemistry, and biology. The majority of the questions in the dataset are presented alongside a supporting paragraph containing the relevant scientific information, thereby enabling reading comprehension-based assessment (Clark et al., 2018; Rajpurkar et al., 2016).

The dataset used in this study consists exclusively of multiple-choice questions (MCQs), each containing one correct answer and three distractors. Therefore, the evaluation system developed in this study is specifically designed for multiple-choice assessment contexts and does not directly evaluate open-ended responses. This constraint should be considered when interpreting the generalizability of the findings.

The dataset contains classified question titles and the correct answer for each question, as well as three carefully constructed 'distractors' (misleading options). Additionally, for some questions, evidence texts determined by human experts are provided to evaluate model performance. This structure enables the evaluation of not only the model's ability to find the correct answer but also its competence in knowledge-based reasoning (Ganesan et al., 2023; Liu et al., 2020).

Within the scope of this study, only questions directly related to science were analyzed. To this end, questions containing specific scientific keywords were filtered from the dataset, and a separate subset was created. This approach highlights the importance of data pre-processing for the accurate learning of contextual concepts in natural language processing models (Rogers et al., 2020; Tenney et al., 2019).

The keywords used are as follows; atom, element, molecule, reaction, energy, heat, electricity, current, cell, DNA, protein, plant, animal, bacterium, matter, physics, chemistry, biology, pressure, density, light, sound, evolution, nature, planet, sun, earth, universe, mass, temperature, magnet, neutron, electron, gravity, speed, force, living being, organism, nucleus, solar system, photosynthesis, electrical, chemical, fuel, gas, liquid, solid, scientist, experiment, laboratory, microscope, telescope, galaxy.
The rationale for this approach is explained below:

- To increase content validity: Questions with scientific accuracy ensure that the model truly assesses science-related concepts (Cohen et al., 2018; Messick, 1995).
- To train the model with conceptual structures specific to the field of science: This has enabled the model to learn specific patterns in the context of science education more clearly (LeCun et al., 2015; Vaswani et al., 2017).
- Ensuring pedagogical alignment: Filtering enabled the selection of content appropriate for curricula and classroom applications (National Research Council [NRC], 2012).
- Creating data homogeneity: By filtering out off-topic questions, the model is trained on cleaner and more consistent data (Beltagy et al., 2019; Brown et al., 2020).

The dataset is divided into three subsets: 'train', 'validation', and 'test'. Since only publicly available data was used, no ethics committee approval was required. The SciQ dataset was shared via a public platform and does not contain any personal data or confidential information. Therefore, only openly accessible, anonymized, and ethically compliant data was used in this study. No human participant data or intervention would require ethical committee approval. This indicates that there is no ethical risk in conducting the research (American Psychological Association [APA], 2020; Zawacki-Richter et al., 2019).

Model

In this study, a classification system was developed using DeBERTa-v3-large, one of the pre-trained LLMs, for the automatic evaluation of multiple-choice science questions. The DeBERTa (Decoding-enhanced BERT with disentangled attention) model stands out for its ability to analyse inter-contextual meaning relationships in greater depth and provides more accurate context analysis than classical BERT architectures (He et al., 2021; Zhou et al., 2022). In this regard, it has the potential to perform well in measurement tools that require careful context analysis, such as multiple-choice questions (Liu et al., 2019).

The pre-trained version of the model was loaded using the open-source HuggingFace Transformers library (Wolf et al., 2020) and then fine-tuned on a training subset of the SciQ dataset consisting solely of examples related to the natural sciences. This method is widely used to improve the model's task-specific performance through transfer learning (Devlin et al., 2019; Raffel et al., 2020).

During the data preprocessing process, each question was evaluated separately with four options (one correct answer and three carefully designed distractors), and these options were appropriately coded as input to the model. This application is based on the 'input pair encoding' method commonly used in evaluating multiple-choice questions (Lan et al., 2020). In addition, the model was designed to learn the semantic differences between the four options for each question, so the effectiveness of attention mechanisms with high semantic discrimination power was evaluated (Clark et al., 2019).

The training process was carried out using the HuggingFace Trainer class and TrainingArguments structure. The model's performance was monitored primarily using the accuracy metric, and the hyperparameter settings were optimized according to this metric. Both the training and validation sets were used in the training process, and the risk of overfitting was controlled using an early stopping criterion (Goodfellow et al., 2016).

This approach offers the possibility of rapid, consistent, and explainable assessment in conceptually dense fields such as science, compared to traditional measurement and assessment methods; it also serves as a model for the practical integration of AI-supported applications in the field of educational technology (Baker & Siemens, 2014; Holmes et al., 2019; Luckin et al., 2016).

Table 2
Fine-Tuning Hyperparameters for DeBERTa-v3 Model Training

Hyperparameter	Value
Learning rate	1e-5
Training/Evaluation batch size	4
Gradient accumulation steps	4
Number of epochs	10
Warm-up ratio	0.1
Learning rate scheduler	Cosine decay
Optimization algorithm	AdamW (weight decay = .01)
Early stopping patience	2 epochs

Note. Hyperparameters were selected to maximize accuracy and prevent overfitting based on the validation set.

As shown in Table 2, the hyperparameters were carefully selected to balance model performance and generalization, with particular attention to preventing overfitting.

Evaluation Process

In this study, fundamental concepts of measurement and assessment have been operationalized in accordance with the principles of educational measurement and assessment.

Validity has been defined as the model's ability to accurately identify the correct answer and has been measured using classification accuracy on the test dataset. This criterion demonstrates the extent to which the model accurately represents the targeted scientific knowledge and reasoning skills.

Reliability has been operationalized as the consistency of the model's justification-based decisions. In this context, a cosine similarity analysis was performed between each answer option and the supporting text, and the degree of alignment between the option with the highest semantic similarity and the correct answer was used as an indicator of response consistency and reliability.

Conceptual understanding has been defined as the extent to which the model captures the semantic alignment between the justifications of its responses and scientific explanations. This dimension has been assessed using a justification-based semantic similarity analysis, revealing the extent to which the model reflects deep scientific reasoning processes beyond surface-level accuracy.

The evaluation process was conducted in line with the principles of validity, reliability, and measurement quality, which are the fundamental dimensions of measurement and evaluation science (Birenbaum & Tatsuoka, 2013). Within this scope, three fundamental research questions were defined:

In this study, the independent variables include the question text, answer options, and supporting paragraphs, while the dependent variables consist of model accuracy and semantic similarity scores. These variables were operationalized to evaluate validity, reliability, and conceptual understanding within the framework of automated assessment.

The research questions were formulated in alignment with the operational definitions of validity, reliability, and conceptual understanding.

RQ1. Construct Validity: To what extent does the AI-based automated evaluation system developed using the SciQ dataset demonstrate validity, as measured by its accuracy in identifying correct answers to science questions? The model's accuracy (correct answer rate) was calculated using the test dataset, and the extent to which the model represents the scientific reasoning skills it aims to measure was

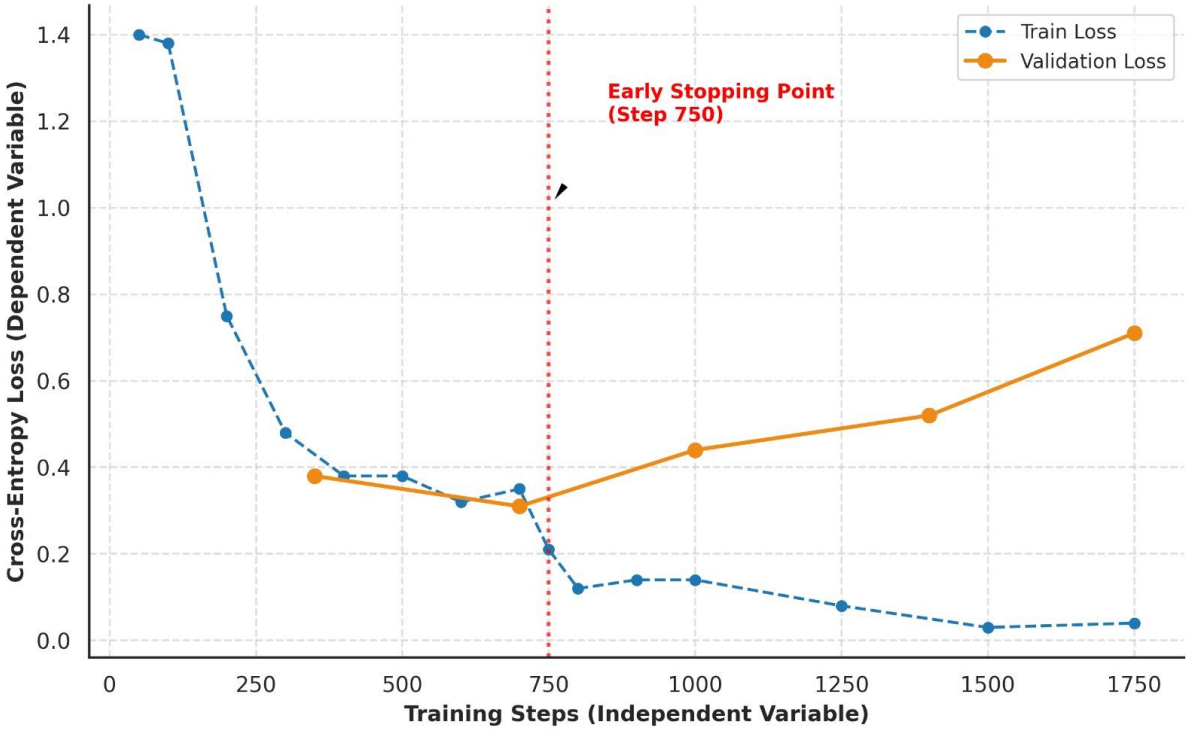
examined. The SciQ dataset covers disciplines commonly used in education, strengthening the model evaluation in terms of content validity (Welbl et al., 2017).

RQ2. Reliability: To what extent does the system demonstrate reliability in evaluating justification-based responses, as measured by the degree of alignment between the most semantically similar option and the correct answer? In the justification-based analysis, cosine similarity between each supporting paragraph and answer option was calculated using the Sentence Transformers (all-MiniLM-L6-v2) model. The degree of alignment between the option with the highest similarity score and the correct answer was examined as an indicator of response consistency and justification competence (Reimers & Gurevych, 2019).

RQ3. Conceptual Understanding and Assessment Quality: To what extent does the system capture conceptual understanding through justification-based semantic alignment within the SciQ dataset, and how does it contribute to the quality of formative and summative assessment practices? By interpreting the findings from RQ1 and RQ2 together, conclusions were drawn regarding the system's applicability in formative (learning process-supportive) and summative (learning outcome-measuring) assessment environments, as well as its potential to facilitate teachers' feedback processes.

Figure 1

Training and Validation Loss Curves Reflecting Learning Dynamics and Overfitting



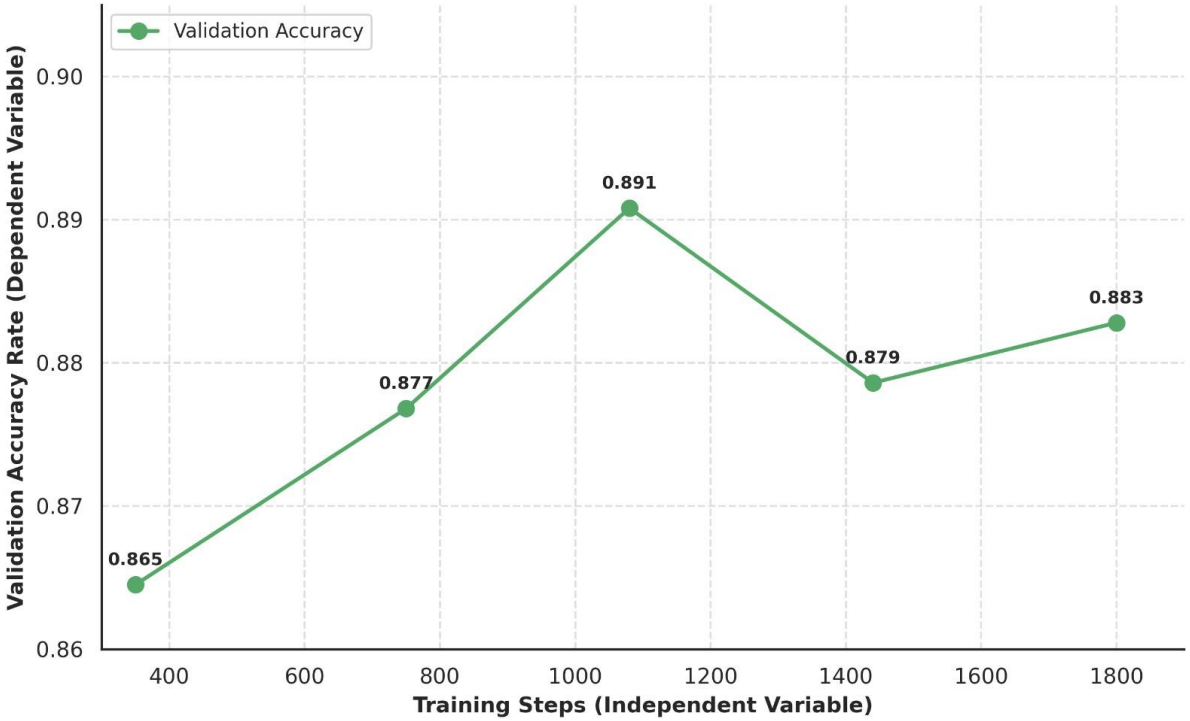
Note. The graph illustrates the relationship between training steps (independent variable) and loss values (dependent variable). The dashed blue line represents training loss, while the solid orange line represents validation loss. The sharp increase in validation loss after approximately 750 steps indicates the onset of overfitting, supporting the application of an early stopping strategy to preserve model generalization.

As shown in Figure 1, training loss decreases significantly as training steps progress. This indicates that the model has successfully learned the patterns in the training data. However, after approximately 750 steps, a noticeable increase in evaluation loss begins. This suggests that the model has lost its generalization ability by overfitting to the training data (Goodfellow et al., 2016). Therefore, applying

the early stopping strategy around the 750th step would be appropriate to maintain the model's overall success rate.

Figure 2

Validation Accuracy Trend of the DeBERTa-v3 Model During Training



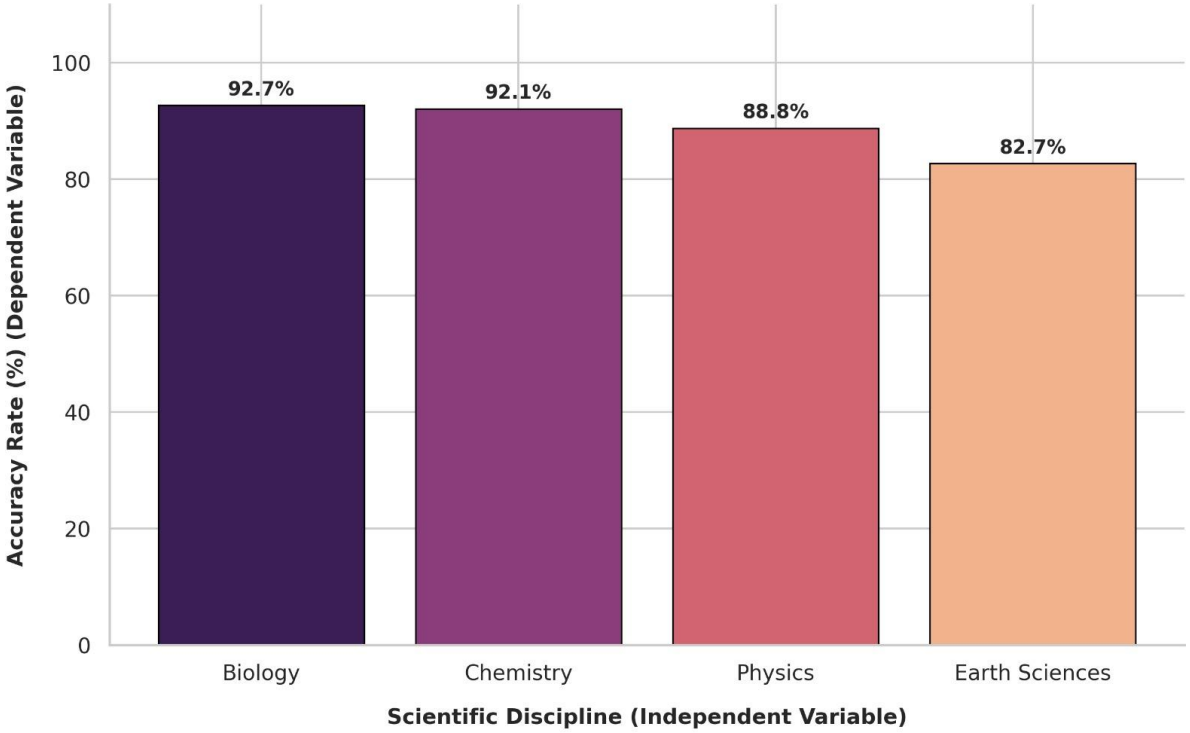
Note. The x-axis represents training steps (independent variable), and the y-axis represents validation accuracy (dependent variable). Accuracy is defined as the proportion of correctly predicted science questions relative to the total number of questions in the validation set. The peak accuracy of 89.1% is observed at approximately step 1080. Subsequent fluctuations suggest the point at which model generalization begins to stabilize.

In Figure 2, the accuracy rate is around 86.5% at the beginning and increases to 89.1% as training progresses. This increase indicates that the model's performance on the validation data has improved. However, a slight decline is observed after the 1000th step, with the model's accuracy dropping to 87.8%; this may indicate signs of overfitting. A slight recovery in accuracy is observed in subsequent steps. This trend highlights the importance of early stopping and regularization strategies in the training process (Goodfellow et al., 2016).

The loss values obtained during the model's training process indicate that the learning process progressed steadily. The training loss decreased regularly throughout each epoch, while the validation loss showed a similar trend. This indicates that the model was able to learn general characteristics without showing a tendency toward overfitting (Goodfellow et al., 2016). In particular, the rapid decrease in both losses in the early epochs shows that the model quickly adapted to the data at the beginning. As the model approaches the final epochs, the losses converge and stabilize, suggesting that the model has learned sufficiently and that early stopping was applied at the appropriate point.

An analysis comparing the success rates of the model in the sub-disciplines of science reveals discipline-based performance differences in the system. These findings are visualized in Figure 3.

Figure 3
Model Performance Across Scientific Disciplines Based on Accuracy Rates



Note. The x-axis represents scientific disciplines (independent variable), and the y-axis represents model accuracy (%) (dependent variable). Accuracy is defined as the percentage of correctly identified answers by the DeBERTa-v3 model within each discipline. The variation in performance—ranging from 92.7% in biology to 82.7% in earth sciences—reflects differences in contextual complexity and terminological density across disciplines.

This situation indicates that the model performs better in distinguishing conceptual questions in some disciplines, while in other areas there is a need for further improvement. In particular, the contextual complexity of earth science questions or the representation ratio in the dataset can be considered among the factors influencing this difference.

A closer examination of incorrect predictions revealed that most errors occurred in cases where distractor options were semantically similar to the correct answer. This indicates that while the model is capable of capturing general contextual meaning, it may struggle in fine-grained semantic discrimination tasks. This finding highlights the importance of improving justification-sensitive evaluation mechanisms.

FINDINGS

Model Performance (RQ1)

The multiple-choice classification model based on DeBERTaV3-large trained on the SciQ dataset achieved an accuracy rate of 90.4% on the test dataset. This result represents a very high level of performance among BERT-derived models used for multiple-choice question answering. In comparisons with similar architectures in the literature, the model is seen to have an advantage due to its context awareness and conceptual disambiguation capabilities (Allard et al., 2024; Khashabi et al., 2020).

At this point, rather than simply presenting accuracy rates, a detailed explanation of the model's successes and failures contributes to a clearer understanding of its strengths and weaknesses (Ganesan et al., 2023). This success demonstrates that the model not only learns word sequences but also

meaningfully analyses the contextual-semantic relationships between the supporting paragraph and the answer options. When examining the confusion matrix analyses, it was determined that the vast majority of examples where the model made errors were due to carefully crafted distractors that were semantically very close to the target answer. This indicates that the model possesses strong conceptual discrimination abilities but occasionally struggles to make decisions between contextually similar options (Devlin et al., 2019).

Reasoning-Based Reliability Analysis (RQ2)

The model was analyzed not only for its ability to select the correct answer, but also for its capacity to provide supporting reasoning for this decision. Such justification-based analyses are pedagogically meaningful because they provide in-depth insights into students' thinking processes beyond simply selecting the correct answer (Khashabi et al., 2020; Shute & Rahimi, 2021). For this purpose, the cosine similarity between each option selected by the model and the supporting paragraph accompanying the question was calculated using the Sentence-BERT architecture. As a result of this analysis, the option with the highest semantic similarity to the supporting paragraph was found to match the correct answer 78.75% of the time.

This ratio shows that the model can make reasoned decisions and base its choices not only on statistical patterns but also on meaning. This also indicates that the model captures aspects of conceptual understanding by aligning responses with underlying scientific explanations. This indicates that AI-based tools can function as systems that not only focus on the outcome but also model students' comprehension processes (Ganesan et al., 2023). Considering that automatic assessment systems used in educational settings must ensure not only accuracy but also pedagogical reliability, the capacity for justification becomes a crucial criterion (Williamson & Eynon, 2020).

Measurement Quality and Potential for Use in Education (RQ3)

Based on the findings obtained within the scope of RQ1 and RQ2, the following conclusions were reached regarding the applicability of the developed model in both summative (total scoring) and formative (process-oriented feedback) assessment processes in the context of science education. This approach emphasizes the need to evaluate not only technical accuracy but also pedagogical meaningfulness, explainability, and qualities that support student learning (Holmes et al., 2019).

- In summative assessment, the model's high accuracy rate appears sufficient for scoring individual achievement and measuring student performance quickly and consistently in exam-like situations. Such applications enable objective and time-efficient assessment, especially in large groups (Maaten et al., 2020).
- In terms of formative assessment, it is possible to relate the model to the level of thinking in student responses through reasoning analysis. The semantic similarity value, which shows the degree to which the student's response matches the supporting text, can provide diagnostic information to the teacher and facilitate the production of targeted feedback (Ganesan et al., 2023; Shute, 2008).
- Advantages: The model offers significant advantages in terms of processing speed, scalability, low human resource requirements, and the ability to generate personalized feedback. In these respects, it has the potential to reduce teachers' assessment workload.
- Limitations: Some distractors that are semantically very close to the target answer can confuse the model. In addition, technical shortcomings such as random variations inherent in language models and limited explanations of reasoning pose obstacles to the model's use in assessment processes in a fully autonomous manner (He & Yih, 2023).

These discipline-based differences indicate that model performance is sensitive to content structure and conceptual complexity across scientific domains, suggesting that dataset composition plays a critical role in automated assessment outcomes.

DISCUSSION

The findings of this study are discussed in relation to the research questions, focusing on validity (RQ1), reliability (RQ2), and conceptual understanding within assessment quality (RQ3).

In relation to RQ1 (validity), the DeBERTa-v3-based model achieved an accuracy rate of 90.4% on the SciQ dataset. This high level of performance indicates that the model can successfully identify scientifically valid answers within a multiple-choice context. Comparable accuracy levels reported in transformer-based models trained on the same dataset (Allard et al., 2024; Khashabi et al., 2020) further support the consistency of these findings with the existing literature. From a measurement perspective, this result provides evidence for the construct validity of the system, as it demonstrates the capacity to represent targeted scientific knowledge and reasoning processes (Kane, 2013; Messick, 1995). However, consistent with prior research, validity in educational assessment should not be reduced to accuracy alone; it also requires meaningful interpretation of learners' responses (Mislevy et al., 2003; Nitko & Brookhart, 2014; Pellegrino et al., 2001).

In relation to RQ2 (reliability), the justification-based semantic similarity analysis showed that the most semantically aligned option corresponded to the correct answer in 78.75% of the cases. This suggests that the model's outputs are not random but reflect a consistent alignment with the underlying scientific context. Previous studies employing SBERT-based approaches similarly report improvements in consistency in automated answer evaluation (Ganesan et al., 2023; Reimers & Gurevych, 2019). Moreover, research on automated analysis of student explanations indicates that such approaches can yield reliable insights into student thinking (Haudek et al., 2011). Taken together, these findings support the use of semantic similarity as a meaningful indicator of reliability in AI-supported assessment systems, while also highlighting that justification-based evaluation remains a complex construct requiring further refinement.

In relation to RQ3 (conceptual understanding and assessment quality), the combined interpretation of accuracy and semantic alignment results suggests that the model captures aspects of conceptual understanding beyond surface-level correctness. By integrating justification-based analysis, the system evaluates not only which answer is selected but also how closely that answer aligns with scientific explanations. This is consistent with prior research demonstrating that machine learning approaches can be used to assess students' conceptual understanding in science (Nehm et al., 2012). In addition, AI-supported systems have been shown to support formative assessment by making students' reasoning processes more visible and by enabling more effective feedback (Holmes et al., 2019; Shute & Rahimi, 2021; Williamson & Eynon, 2020). In this sense, the proposed system shows potential to contribute to both formative and summative assessment practices.

From a pedagogical perspective, the system offers important opportunities to transform classroom assessment practices by shifting the focus from answer-based evaluation to reasoning-based assessment. This perspective aligns with research highlighting the role of explainable AI in enhancing student engagement and promoting deeper learning (Holmes et al., 2019; Luckin et al., 2016). Furthermore, feedback-centered approaches are widely recognized as a key factor in improving learning outcomes (Hattie & Timperley, 2007). Within this framework, the integration of semantic similarity analysis enables teachers to examine not only whether students reach correct answers but also how their reasoning aligns with scientific knowledge, thereby supporting more targeted and meaningful feedback processes.

At the same time, the findings point to important trade-offs between different dimensions of assessment quality. While some models achieve high accuracy, they may offer limited explainability, whereas more interpretable models may perform less accurately. This tension is well documented in the literature (Doshi-Velez & Kim, 2017; Rudin, 2019). Consistent with this, the results of the present study suggest that accuracy, reliability, and explainability should be considered as complementary rather than interchangeable dimensions.

The error analysis further supports this interpretation. The model was found to have difficulty distinguishing between semantically similar distractors, indicating limitations in fine-grained conceptual discrimination. Similar challenges have been reported in prior research on question answering systems (Devlin et al., 2019; Ganesan et al., 2023; Sugawara et al., 2018). This suggests that, although the model captures general contextual meaning effectively, further improvements are needed to strengthen its sensitivity to nuanced reasoning differences.

In addition, the characteristics of the dataset play a crucial role in shaping model performance. While the SciQ dataset offers strong content validity due to its structured and curriculum-aligned design (Clark et al., 2018; Welbl et al., 2017), the limited realism of some distractors may constrain the assessment of deeper reasoning processes. This observation is consistent with research emphasizing the importance of data quality in AI-based systems (Geiger et al., 2020; Zawacki-Richter et al., 2019).

Finally, these findings should be considered in light of broader ethical and pedagogical considerations. The effective use of AI in education requires not only technical performance but also alignment with principles such as transparency, fairness, and pedagogical relevance (Holmes et al., 2019; Kizilcec, 2022). Ethical frameworks further highlight the responsibility of AI systems to support equitable and accountable educational practices (Floridi et al., 2018). Therefore, AI-supported assessment systems should be understood as tools that complement, rather than replace, teacher judgment.

Overall, the findings of this study reinforce existing literature emphasizing that AI-based assessment systems should be evaluated not only in terms of accuracy but also with respect to validity, reliability, explainability, and pedagogical effectiveness (Holmes et al., 2019; Williamson & Eynon, 2020). This underscores the importance of integrating educational theory into the design and evaluation of AI-supported assessment tools.

CONCLUSION

In this study, the AI-supported assessment model developed with the DeBERTa-v3 architecture based on the SciQ dataset demonstrated its potential as a valid and partially pedagogically reliable assessment tool, offering high accuracy, meaningful justification, and limited explainability.

One limitation of this study is the reliance on the SciQ dataset, which may not fully represent real classroom conditions. Future studies should validate the model using real student data to improve ecological validity.

The model can be used as a supportive tool to reduce teachers' assessment burden, particularly with multiple-choice questions, to quickly assess students' conceptual understanding, and to provide individualized feedback. However, rather than eliminating teacher interpretation, it is recommended that such systems be integrated into blended assessment processes (Luckin et al., 2016).
The following recommendations could be considered in future studies:

- Tests for language and culture dependency should be conducted. Since the SciQ dataset is in English, the validity of the model should be re-evaluated with datasets specific to Turkish and different cultural contexts.
- Model outputs should be visualized to increase explainability. Methods such as heatmaps, attention weights, or justification explanations can be used to facilitate teachers' understanding of model decisions (Ganesan et al., 2023).
- Model outputs should be linked to student thinking processes. This allows for an analysis of how students think, not just whether they know the correct answer.
- AI-based assessments should be integrated with teacher observations. Comparative analysis of qualitative data (e.g., student explanations) and quantitative model outputs can provide a more holistic assessment process.

In conclusion, this study demonstrates the potential of AI-supported assessment systems in science education, but it also clearly identifies their pedagogical limitations. Therefore, integrating technical advancements into the educational context is critical for the responsible and meaningful use of AI in learning ecosystems. This study demonstrates that AI-supported assessment systems can move beyond answer evaluation toward reasoning-based assessment, offering a significant advancement in the quality of science education assessment practices.

REFERENCES

Allard, A., Smith, J., & Zhao, L. (2024). Enhancing AI reasoning through chain-of-thought prompting: A study using AQuA-RAT dataset. *Journal of Educational Artificial Intelligence, 7*(1), 45–63. https://doi.org/10.48550/arXiv.2305.16582

American Educational Research Association. (2014). *Standards for educational and psychological testing* (2nd ed.). AERA.

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). APA.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Baker, R. S., & Siemens, G. (2014). Educational data mining and learning analytics. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 253–272). Cambridge University Press.

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics, 7,* 49–72. https://doi.org/10.1162/tacl_a_00254

Beltagy, I., Lo, K., & Cohan, A. (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)* (pp. 3615–3620). https://doi.org/10.18653/v1/D19-1371

Birenbaum, M., & Tatsuoka, K. K. (2013). *Assessment for learning: Building a sustainable assessment culture*. Routledge.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Borji, A. (2023). *A categorical archive of ChatGPT failures*. arXiv. https://arxiv.org/abs/2302.03494

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading.* ASCD.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). *Language models are few-shot learners.* arXiv. https://arxiv.org/abs/2005.14165

Bull, S., & Kay, J. (2016). SMILI: A framework for interfaces to learning data in open learner models. *International Journal of Artificial Intelligence in Education, 26*(1), 293–331. https://doi.org/10.1007/s40593-015-0092-1

Chen, C. M., Xie, H., & Hwang, G. J. (2020). A multi-faceted review on the applications of artificial intelligence for personalized adaptive learning. *IEEE Access, 8*, 180457–180473. https://doi.org/10.1109/ACCESS.2020.3026097

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). *What does BERT look at? An analysis of BERT's attention*. arXiv. https://arxiv.org/abs/1906.04341

Clark, C., Tafjord, O., & Mishra, S. (2018). *Think you have solved question answering? Try ARC, the AI2 reasoning challenge.* arXiv. https://arxiv.org/abs/1803.05457

Cohen, R. J., Swerdlik, M. E., & Sturman, E. D. (2018). *Psychological testing and assessment: An introduction to tests and measurement* (9th ed.). McGraw-Hill Education.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*

(Volume 1: Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning.* arXiv. https://arxiv.org/abs/1702.08608

Farrow, R. (2023). The possibilities and limits of explicable artificial intelligence (XAI) in education: A socio-technical perspective. *Learning, Media and Technology, 48*(2), 266–279. https://doi.org/10.1080/17439884.2023.2185630

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines, 28,* 689–707. https://doi.org/10.1007/s11023-018-9482-5

Furtak, E. M., Morrison, D., & Kroog, H. (2016). Investigating the link between learning progressions and student responses. *Science Education, 100*(5), 902–928. https://doi.org/10.1002/sce.21229

Ganesan, A., Liu, R., & Koedinger, K. (2023). Justification-based automatic answer evaluation using SBERT. In *Proceedings of the 16th International Conference on Educational Data Mining (EDM 2023)* (pp. 315–324). International Educational Data Mining Society.

Geiger, R. S., Cope, D., Ip, J., Lotosh, M., Shah, A., Weng, J., & Tang, R. (2021). "Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? *Quantitative Science Studies, 2*(3), 795–827. https://doi.org/10.1162/qss_a_00144

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2011). What are they thinking? Automated analysis of student writing about genetics. *CBE—Life Sciences Education, 10*(1), 64–76. https://doi.org/10.1187/cbe.11-08-0084

He, P., Liu, X., Gao, J., & Chen, W. (2021). *DeBERTa: Decoding-enhanced BERT with disentangled attention. In International Conference on Learning Representations.* OpenReview. https://openreview.net/forum?id=XPZIaotutsD

He, X., & Yih, W. T. (2023). Explaining AI predictions in education: From black-box to glass-box. *Educational Technology Research and Development, 71,* 333–349. https://doi.org/10.1007/s11423-022-10156-3

Heffernan, N., & Heffernan, C. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education, 24*(4), 470–497. https://doi.org/10.1007/s40593-014-0024-x

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. Corwin Press.

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign. http://udaeducation.com/wp-content/uploads/2019/05/AIED-Report-2019.pdf

Jurafsky, D., & Martin, J. H. (2020). *Speech and language processing* (3rd ed., draft). https://web.stanford.edu/~jurafsky/slp3/

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kay, J., Bull, S., Gibson, A., & Heffernan, N. (2022). Explainable AI for education: Theory, practice and critique. *British Journal of Educational Technology, 53*(4), 783–798. https://doi.org/10.1111/bjet.13144

Khashabi, D., Min, S., Khot, T., Sabharwal, A., & Hajishirzi, H. (2020). *UnifiedQA: Crossing format boundaries with a single QA system*. arXiv. https://arxiv.org/abs/2005.00700

Kizilcec, R. F. (2022). Algorithmic fairness in education. *Computers and Education: Artificial Intelligence, 3*, 100052. https://doi.org/10.1016/j.caeai.2022.100052

Lehrer, R., & Schauble, L. (2006). Scientific thinking and science literacy. In K. A. Renninger & I. E. Sigel (Eds.), *Handbook of child psychology* (6th ed., Vol. 4, pp. 153–196). Wiley.

Lai, E. R., & Viering, M. (2012). *Assessing 21st century skills: Integrating research findings*. Pearson.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv. https://arxiv.org/abs/1907.11692

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education.* Pearson Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning. *Journal of Research in Science Teaching, 49*(7), 907–935. https://doi.org/10.1002/tea.21029

Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessment of students* (7th ed.). Pearson.

Osborne, J. (2014). Teaching scientific practices: Meeting the challenge of change. *Journal of Science Teacher Education, 25*(2), 177–196. https://doi.org/10.1007/s10972-014-9384-1

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press.

Popham, W. J. (2017). *Classroom assessment: What teachers need to know* (7th ed.). Pearson.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67. http://jmlr.org/papers/v21/20-074.html

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2383–2392). https://doi.org/10.18653/v1/D16-1264

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. arXiv. https://arxiv.org/abs/1908.10084

Rudin, C. (2019). Stop explaining black box models for high stakes decisions. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Shavelson, R. J. (2006). On the integration of formative assessment in teaching and learning. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 453–468). Lawrence Erlbaum Associates.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Shute, V. J., & Rahimi, S. (2021). Review of modern psychometrics applied to learning analytics. *Journal of Learning Analytics, 8*(1), 6–20. https://doi.org/10.18608/jla.2021.7096

Sugawara, S., Kido, Y., Yokono, H., & Aizawa, A. (2018). Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)* (pp. 4093–4104).

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Wang, M.-T., & Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review, 29*(1), 119–140. https://doi.org/10.1007/s10648-015-9355-x

Welbl, J., Liu, N. F., & Gardner, M. (2017). Crowdsourcing multiple-choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text (W-NUT 2017)* (pp. 94–106). https://doi.org/10.18653/v1/W17-4413

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv. https://arxiv.org/abs/2201.11903

Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology, 45*(3), 223–235. https://doi.org/10.1080/17439884.2020.1798995

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). https://doi.org/10.18653/v1/2020.emnlp-demos.6

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), Article 39. https://doi.org/10.1186/s41239-019-0171-0

Zhou, C., Li, W., Shen, S., Lu, D., & Zhang, H. (2022). On the explainability of transformers in multiple-choice QA. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4635–4647). https://doi.org/10.18653/v1/2022.emnlp-main.317