JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

JSCDM

Vol. 6 No. 3 (2025) 323-355
https://publisher.uthm.edu.my/ojs/index.php/jscdm

The Use of Transformer-Based Models for Automatic Short-answer Scoring in Education: A Systematic Literature Review

Widhy Hayuhardhika Nugraha Putra[1,2]*, Mohd Farhan Md Fudzee[1,2], Buce Trias Hanggara[1,3], Welly Purnomo[1], Fajar Pradana[1], Admaja Dwi Herlambang[1]

[1] *Brawijaya University, Faculty of Computer Science, Lowokwaru, Malang, 65145, INDONESIA*
[2] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, 86400, MALAYSIA*
[3] *PT. Garapan Teknologi Indonesia (GTI – ICT industry), Lowokwaru, Malang, 65145, INDONESIA*

*Corresponding Author: widhy@ub.ac.id
DOI: https://doi.org/10.30880/jscdm.2025.06.03.021

Article Info

Received: 29 October 2025
Accepted: 14 December 2025
Available online: 30 December 2025

Keywords

Automatic short-answer scoring, quality education, large language model, AI for education

Abstract

This review focuses on recent advancements in the Automatic Short-Answer Scoring (ASAS) system in education. The primary objective of this review is to identify current trends in utilizing transformer-based models for the ASAS system. This review also aims to discuss future directions for ASAS technology. ASAS's conventional machine learning methods were inconsistent because they rely on statistical similarity and are prone to bias. Meanwhile, transformer-based models were typically used for feature extraction, embedding, and score calculation via classification or regression. They generally served as a similarity calculator comparing students' answers to the reference answer. We applied the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework to explore ASAS's state of the art and uncover its future trends. Our findings reveal that transformer-based models significantly outperform traditional machine learning approaches by capturing complex context. On the other hand, LLMs excel at providing feedback and score justification. Recent studies have shown a shift toward using transformer-based models for ASAS's complex tasks, including data augmentation and feedback generation. However, further research is needed to use LLMs and GPTs to generate explainable, fairer scores and to address data scarcity through reasoning and augmentation.

1. Introduction

Transformer-based language models have brought significant improvements to Natural Language Processing (NLP) tasks, particularly for tasks that require context understanding. Short-answer scoring is one such task that transformer-based models have utilized [1]-[5]. The challenge in the short-answer scoring task is accurately grading open-ended responses, given the limited text resources. Traditional scoring methods depend on human evaluators, which can be slow and subjective. Transformer-based models offer a promising alternative because they can capture semantic relationships and context details [6]-[9]. This study investigates the application of

transformer-based models for an Automatic Short-Answer Scoring (ASAS) system [10]-[13]. We aim to identify current trends in transformer-based models for ASAS and contribute to the discussion on the technological advancements in Artificial Intelligence for educational assessments.

Short-answer assessment responses were shorter than the essay assessment responses. Short-answer assessments often elicit more varied responses within a limited number of words [14]-[17]. In contrast, essay scoring typically involves longer, more structured texts that focus on coherence, grammar, and style [10], [18]. These differences would affect the way the score is calculated. The most used answer assessment method is manual, which involves human assessment. Teachers must read the entire essay and grade it individually, which is time-consuming and potentially subjective.

As short answers often provide limited text, a deeper understanding of the context is important in a short-answer scoring system. In a short-answer assessment, students were required to express their knowledge and encapsulate their complex ideas within a constrained word count. Consequently, ASAS models were challenged to accurately assess students' understanding of the course materials from their limited answers [19]-[22]. Therefore, the context around a short-answer assessment is crucial for fairly assessing the quality and relevance of the answers. By training models on varied linguistic styles and expressions, researchers try to mitigate biases that may arise. Incorporating techniques such as knowledge distillation, additionally, could gain more model performance without compromising accuracy [23, 24].

As the landscape of NLP continues to evolve, advancements in automated scoring systems enhance system efficiency and performance [25]-[28]. It also raises critical questions regarding equity and accessibility in assessments. Although transformer models show remarkable accuracy in evaluating student responses, their effectiveness can be influenced by factors such as language proficiency and diverse expression styles among students [14, 19, 24]. This phenomenon needs careful consideration of how these technologies are deployed across various educational contexts to ensure that they do not disappoint learners. Furthermore, ongoing research that seeks to adapt these models to specific fields could provide insights that improve the ASAS system's flexibility and fairness [29]-[33].

Our review explored existing studies on and ASAS systems. We reviewed recent research to gain a global understanding of the ASAS landscapes. The "Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)" framework was employed to ensure a comprehensive and transparent process [34]. The PRISMA framework provides a structured approach to identify, select, and critically filter the studies. The chosen studies were reviewed and summarized to provide an evidence-based overview of the current research on ASAS.

This study has four research questions. The first question is **"**What are the recent architectures of the automated short-answer scoring systems?" (RQ1). The second question is "How is the transformer-based model used for an automatic short-answer scoring system architecture?" (RQ2). The third question is "What are the general and context-specific datasets available for automatic short-answer scoring model training?" (RQ3). Moreover, the last question is "What recent frameworks have been developed for automatic short-answer scoring?" (RQ4). These questions become our direction of study.

Initially, we defined the study's objectives, which clearly identify the key research themes of our study. Next, we created a clear search strategy through the Scholar databases. We applied strict inclusion and exclusion criteria to choose studies that aligned with our research objectives. We also made a solid quality assessment checklist to ensure the reliability of the selected studies. Finally, we conducted a detailed analysis of the reviewed papers to extract essential insights and findings to help us understand the ASAS fields.

Fig. 1 provides a visual representation of the structure and organization of this study. The rest of this paper is organized into 9 sections. Section 1 describes the introduction and background of this study. Section 2 provides an overview of related work on ASAS systems. Section 3 outlines the methodology of this study and explains how we selected the literature. Sections 4, 5, 6, and 7 present the results of the literature analysis and describe the technological approaches of ASAS. Section 9 discusses the considerations arising from the ASAS study, results, and outlines the limitations of the review.

J. of Soft Computing and Data Mining Vol. 6 No. 3 (2025) p. 323-355

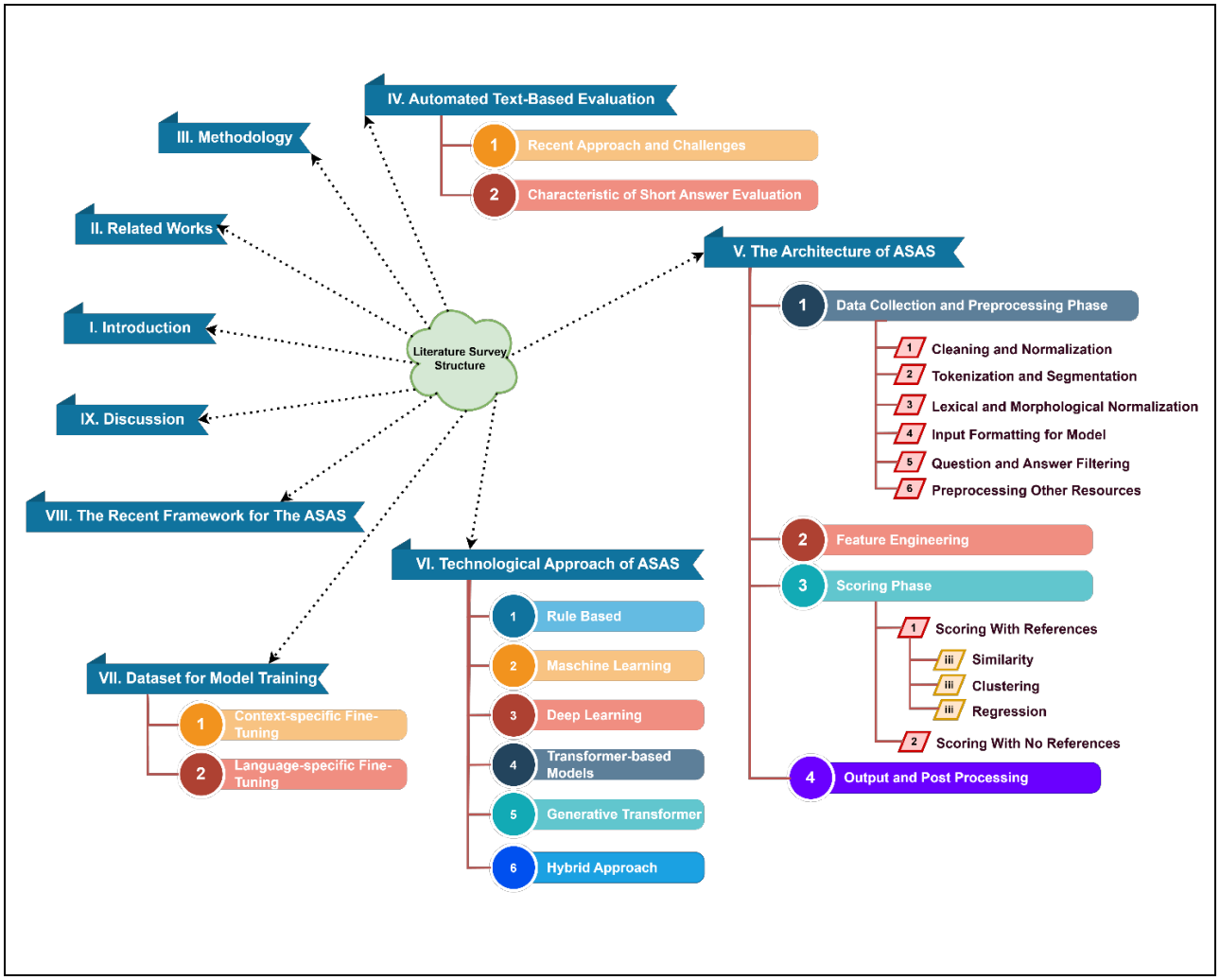

Fig. 1 *Paper organization and structure*

2. Related Works

Researchers have rapidly explored the field of text-based evaluation techniques. For example, a systematic literature review was conducted that focuses on the automatic grading of open-ended written assignments [27], [35]-[38]. Their research explored trends and techniques for Automatic Essay Scoring (AES). The other studies analyzed automatic scoring and feedback in the educational context [39, 40]. They emphasize the need for improvements to mitigate the adverse effects and enhance educational quality and student experience. The gap between short answers in particular domains and long essays also makes this research field challenging and requires sophisticated methodologies [41]-[43].

Researchers have also explored the advantages of an automatic scoring system for education. An automatic scoring system for education enables accommodating larger student numbers without increasing the instructor's workload [41]-[45]. Students' learning experience was enhanced by automatic scoring systems through quicker feedback and reduced score bias [46]-[48]. However, the automatic scoring systems also highlight drawbacks, including the inability to capture innovative students' answers and potentially train students to focus more on answering questions rather than understanding concepts.

A study examined the "strengths and weaknesses of automated scoring of free-text student answers" [49]. They examined various models for automated content scoring and compared their performance, architectures, efficiencies, and transferabilities. Unlike humans, who may be impacted by fatigue or mood, a model consistently assigns the same score to identical answers [50]-[53]. However, models faced a significant challenge in explaining how they arrive at their scores. Deep learning-based models often perform well with minimal manual input. Their black-box nature, however, makes it hard to understand how a score is concluded [54]-[56]. Thus, the topic that needs further exploration is personalized feedback, which makes the score more explainable [57]-[59].

Another study reviewed "Transformer-Based Pre-Trained Language models (PLMs) that use a self-supervised method of learning" [60]. The components of the transformer architecture, such as the encoder, decoder, and

attention mechanisms, were explored. The study also explained the concept of self-supervised learning in a model and how to apply it to massive, unlabeled datasets. Specific PLMs (including BERT, RoBERTa, ELECTRA, and T5) and their proposed pre-training techniques were compared and discussed. They used several datasets, which were categorized into three groups: general-purpose, domain-specific, and multilingual.

The importance of context in the ASAS system was discussed in research [61]. They found that the main challenge for the ASAS is the lack of relevant training data in the domain. These ASAS problems can be addressed using various approaches, which can be broadly classified into two categories: traditional methods that rely on handcrafted features and deep learning approaches. Other researchers discussed various supervised machine-learning techniques for both free-text evaluation [12]. They grouped these machine learning techniques into three categories: feature-based, neural, and hybrid approaches. They discovered that, unlike AES, which evaluates writing skills such as style and structure, ASAS assesses the accuracy of students' answers. Additionally, AES is often used in language classes, whereas ASAS is more frequently applied in subjects such as mathematics and science. Two groups of ASAS studies were also concluded: ASAS without reference answers and ASAS with reference answers. This is similar to research [49]which grouped the ASAS into instance- and similarity-based.

Other researchers evaluated an embedding approach for developing models for ASAS systems [62]. They reviewed the literature published between 2016 and 2021. They identified relevant articles that used embedding to represent the short-answer grading tasks. Generally, these approaches are categorized into non-learning and learning-based approaches. The scoping review highlights the contribution of embeddings to the performance of ASAS models. Several embedding methods were discussed in their research to identify the advantages and disadvantages of each and to examine how embedding is used. The embedding techniques discussed are: Word, Contextual, Sentence, and Sense embedding. They summarized that word embeddings were the most commonly used, followed by contextual and sentence embeddings, whereas sense embeddings were underrepresented. The use of contextual and sentence embeddings is less effective than that of word embeddings, and there remain potential research areas. Embeddings are then used to process ASAS by measuring similarity or are directly used to perform score classification. On a small scale, embedding was also used directly or to initiate long-short-term memory (LSTM). However, their study did not explicitly discuss how the embedding technique was implemented in the various ASAS frameworks.

A review of AES systems for education scenarios was conducted in [63]. They uncover current AES models and AI technologies, including machine learning, deep learning, and natural language processing techniques. The statistical, linguistic, content, and contextual features extracted from the essays are also examined. The datasets and evaluation metrics used, such as Quadratic Weighted Kappa, Mean Squared Error, and Pearson Correlation Coefficient, were reviewed. Their research identified potential areas for future work: enhancing AES to handle diverse essay topics and writing styles, improving model accuracy, ensuring the AES system's reliability and consistency, handling malicious or adversarial essays, and expanding its capabilities. Their study, however, is limited to published articles between 2018 and 2023.

The AES systems and the resulting practical insights for educators were discussed in research [64]. Thus, the development of effective AES models that balance accuracy, fairness, and generalizability was suggested. Additionally, nine prominent AES methods were outlined, which include five prompt-specific and four cross-prompt approaches. The prompt-specific methods are: Support Vector Machine (SVM), which uses a comprehensive set of carefully engineered features to train an SVM-based scoring model; SKIPFLOW-LSTM, which incorporates coherence features related to the semantic similarity between sentences using a neural tensor layer and LSTM; CNN-LSTM-ATT, which employs a hierarchical neural network architecture with CNN for word relations, LSTM for sentence relations, and an attention mechanism.

Researchers [8] conducted a study to investigate the fairness of ASAS's algorithm. The study focused on demographic disparities related to gender and language. The research employed various machine learning methods and found no evidence of gender bias. However, the research identified a slight bias against native language speakers. Important factors affecting fairness, as concluded from their study, include scoring accuracy, language differences, and the distribution of student performance and item difficulty. Their study also suggests approaches to reduce bias and acknowledge limitations related to data and the measurement of language use.

Other researchers [53] conducted a study to compare machine learning and deep learning models with LLMs for the ASAS system in English and Portuguese. Models such as SVM, Random Forest, and BERT were compared with prompt engineering for GPT-4o. Their study used the Texas and "PT_ASAG" datasets in Portuguese and concluded that prompt engineering elements significantly affected performance. The few-shot examples were crucial for English, while detailed reasoning was crucial for Portuguese. A notable finding from their research was that traditional models generally outperformed LLMs in the ASAS system.

In this study, we aim to delve deeper into studies that enhance the ASAS method, focusing on the base models used, the context and language, and the performance tests conducted. This review comprehensively explains the role of transformer-based pre-trained language models in natural language processing applications. Finally, this paper highlights a potential research area of NLP that can be implemented in ASAS [65]. A comparison of this review with related works is presented in Table 1.

Table 1 *Comparison of this study with other reviews*

Reference, Year	Type of Text	Technological Approach				DS	FW	ET	Remarks
		TA	ML	TR	Gen-AI				
[27], 2021	ASAS	×	▲	▲	×	×	×	×	Focused study on current trends and techniques in the automatic grading of open-ended written assignments.
[66], 2021	ASAS & AES	▲	▲	×	×	×	▲	×	Focused study of the student experience of using automatic feedback and automatic scoring
[20], 2021	ASAS	×	√	×	×	√	▲	▲	Focused on comparing various Deep Learning models for ASAS
[12], 2023	ASAS & AES	×	▲	▲	×	▲	√	√	Focused on current machine learning techniques for both short and essay evaluation in education
[62], 2023	ASAS	×	▲	√	×	▲	√	×	Scoping review that summarizes relevant literature on the use of embeddings in automatic short answer scoring (ASAS)
[17], 2024	ASAS	▲	▲	▲	×	×	×	×	Focused on determining the ASAS model classification
This paper	ASAS	√	√	√	√	√	√	√	

√: indicates the topic is well covered, ×: is uncovered, and ▲is partially covered, TA: traditional approaches including rule-based grading and feature comparison, ML: machine learning, TR: transformer-based model, GEN-AI: generative AI model, DS: dataset, FW: future works, ET: evaluation and tools

3. Methodology

To ensure a rigorous and transparent process, this review follows the "Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)" framework [34]. The framework provides a comprehensive and transparent review to facilitate the evaluation of the findings. For an efficient PRISMA process, the findings from the studies were carefully recorded in RIS format using Mendeley reference management software. Mendeley was used to manage references and avoid overlaps between the papers that we collected from different databases.

3.1 Database Selection

The literature search was conducted across five major academic databases: ScienceDirect, Springer Nature Link, Scopus, Web of Science, and IEEE Xplore. These databases were chosen for their wide coverage of high-quality research articles in the fields of computer science, education, and artificial intelligence.

3.2 Search Strategy

This search strategy was designed to capture relevant studies on ASAS. The search keywords were "automatic short-answer scoring" and "automatic short-answer grading. The search query used these keywords and defined parameters, including full-text availability and publications from 2019 to 2025. By integrating these keywords into the search query, studies that directly intersect short-answer grading within a designated time frame were uncovered. Based on the database search, we describe the articles obtained by year and journal or conference type in Fig. 2. We found that this topic will continue to increase in 2024-2025.

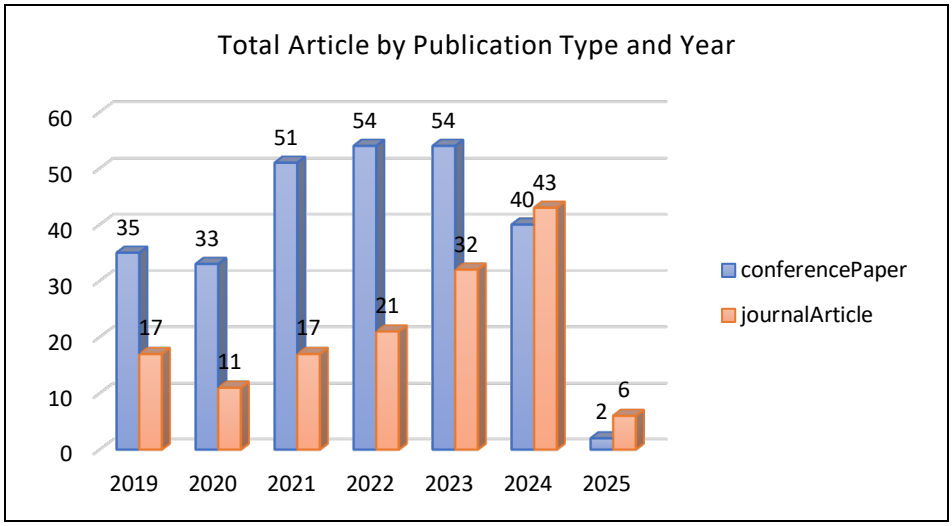

Fig. 2 *Total article by publication type and year*

3.3 Inclusion and Exclusion Criteria

In this review, specific inclusion and exclusion criteria were defined to ensure the selection of relevant and high-quality research articles. By applying these inclusion and exclusion criteria, the systematic literature review aimed to select relevant, recent, peer-reviewed, and empirical research articles on the short-answer grading problem that employ NLP techniques. The inclusion and exclusion criteria for this review are summarized in Table 2.

- **Inclusion Criteria.** Studies were included if they focused on developing, implementing, improving, or evaluating NLP methods for ASAS tasks. Articles discussing language-specific or content/domain-specific trained models were sought after.
- **Exclusion Criteria.** Studies were excluded if they were not peer-reviewed, were not written in English, or did not provide sufficient methodological details. Papers not focusing on NLP tasks for ASAS or those that were purely theoretical without empirical tests were also removed. Articles that focused on question or answer generation were also excluded.

Table 2 *Inclusion and exclusion criteria of the study*

Criteria	Inclusion	Exclusion
Publication Period	January 2019 - February 2025	Outside the inclusion period
Publication Type	Journal articles, conference articles	Masters/Ph.D. Theses, book chapters, abstracts, preprints only, not yet published articles
Study focus	Studies discussing Automatic Short Answer Grading/Scoring	Studies not focusing on Automatic Short Answer Grading/Scoring
Domain Area	Studies discussing Short-Text related evaluation	Studies discussing Long-Text evaluation

3.4 Study Selection

Titles and abstracts were screened to remove irrelevant studies. The remaining articles were then reviewed to ensure that they met the inclusion criteria. Any discrepancies in the study selection were resolved through discussion among the reviewers. To ensure the quality of the included study, we carefully assessed its quality. The Quality Assessment checklists that we use are:

- Are the research questions or research objectives of the paper clear?
- Is the research paper focusing on the scoring method?
- Does the research paper utilize only text-based assessment as a dataset?

Based on the results of the Quality Assessment, we found that 49 papers could not meet all of the criteria we provided. Thus, the remaining papers are all involved in the next stage.

3.5 Data Extraction and Synthesis

Filtered studies were extracted using a standardized form. Key information, including study objectives, dataset, trained models, language, results, and conclusions, was recorded. We then categorized the studies into method-specific, language-specific, and domain-specific categories. The extracted and categorized data from the studies were then synthesized to provide a comprehensive overview of the current state of ASAS, focusing on method-specific, language-specific, and content/domain-specific trained models.

The overall PRISMA flow diagram is shown in Fig. 3. This diagram provides a visual representation of the study selection process. This diagram also notes the number of records identified, included, and excluded, along with the reasons for exclusion. The search yielded a number of relevant articles that were observed in this study.

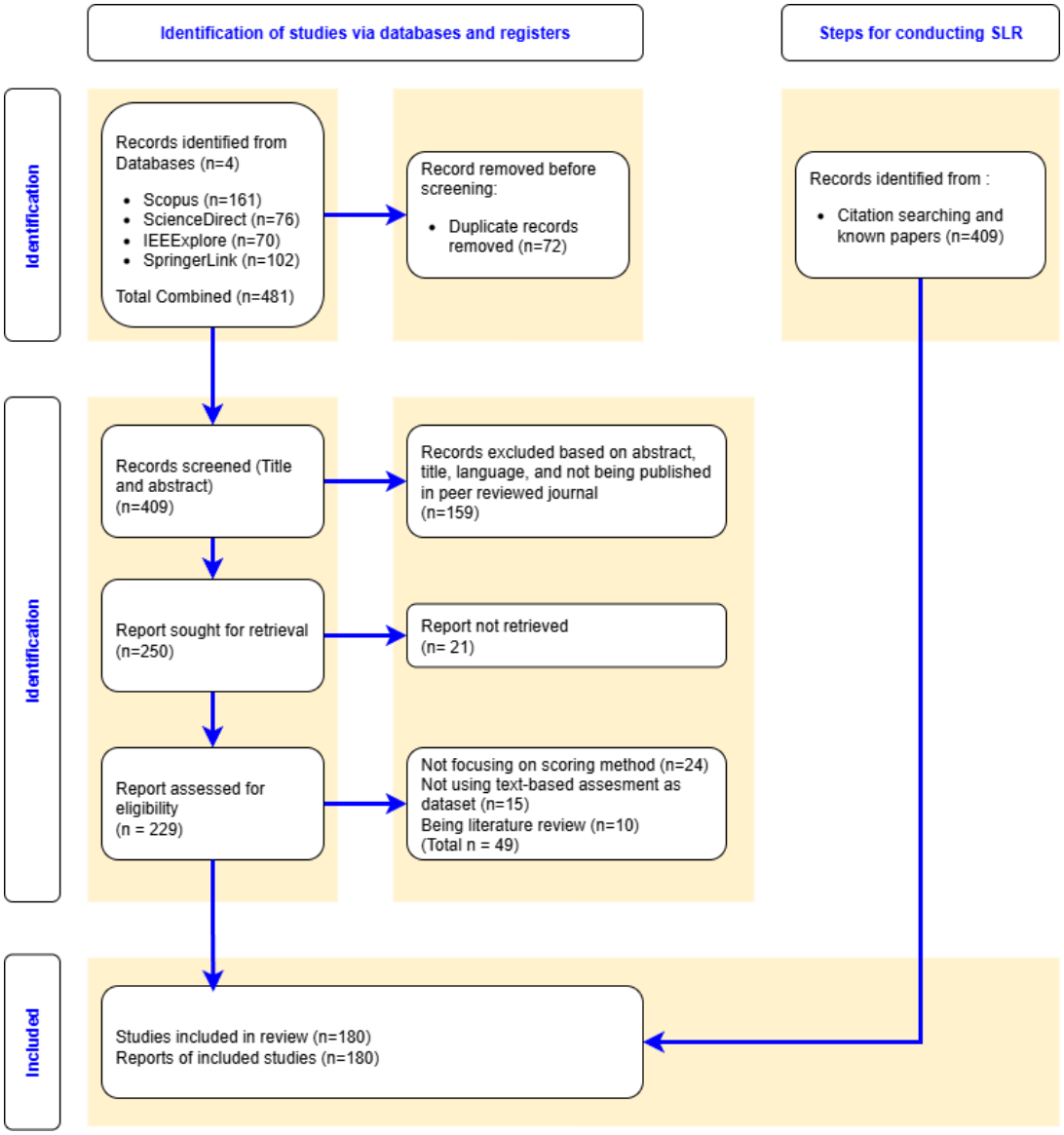

Fig. 3 *Article retrieval and selection process based on the PRISMA reporting standard* [34]

4. Automated Text-Based Evaluation

In this chapter, we discuss automated text-based evaluation in general, including technological approaches currently used and the challenges faced. We also identified several fundamental characteristics of short answers compared with essays, which will be the basis for our next in-depth studies.

4.1 Recent Approach and Challenges

Various approaches and challenges have encountered in automated text-based evaluation tasks [67]-[89]. The critical challenges for automated text-based evaluation tasks were data sparsity, ethical considerations [65, 66], reliability, and time [18]. To address these challenges, researchers have developed various NLP approaches. Feature extraction and text encoding were early methods used to obtain the text's lexical vector. Feature extraction methods extract features from text, like length-based features, lexical features, category-based features, and semantic features [67]-[81]. Attention mechanisms and word-embedding methods were also used to provide a better understanding of the context [82]. Several researchers used self-trained word embeddings and pooling operations to score essays directly [90]-[92].

Advanced transformer-based language models, such as BERT, RoBERTa, DistilBERT, and ALBERT, have been fine-tuned to address several problems in automated text-based evaluation tasks [93]-[123]. These models have the capabilities to extract deep semantic features. With semantic features, models could understand the context and improve scoring accuracy. Research [72] have demonstrated the language models' ability to capture contextual relations between words and sentences. Their study also showed that the models' accuracy improved when combined with other models. Contextual relations also successfully enhanced the models' understanding of content and quality [55, 88].

4.2 Characteristics of Short-Answer Evaluation

A research [124] mentioned that short answers typically consist of two or three sentences in length, and ASAS requires domain-specific knowledge to evaluate and score students' short answers accurately. Short-answer questions allow students to provide their own open-ended text responses rather than selecting from multiple-choice options [106]. Answering short-answer questions requires students to recall information from their knowledge and provide reasons or justifications for their answers. Short-answer responses are typically scored using holistic rubrics with score points, such as 0, 1, 2, etc., based on the quality of the answer. According to [124], the main character distinctions between essay and short-answer scoring are shown in Table 3.

Table 3 *Characteristic comparison between essay scoring and short answer scoring*

	Essay Scoring	*Short Answer Scoring*
Nature of Response	Essays are longer and require evaluation of structure, coherence, grammar, and style. The focus is on the overall quality of writing and argumentation.	Short answers are concise and require a focus on content accuracy and relevance. The challenge lies in understanding the semantic content and comparing it to reference answers.
Evaluation Techniques	Utilizes techniques that analyse linguistic features such as syntax, vocabulary, and organization.	Employs methods like graph convolutional networks and multiway attention networks to craft a model that calculates semantic relations between students' and teachers' answers.

5. The Architecture of ASAS (RQ1)

Researchers have engineered various ASAS architectures to achieve optimal accuracy. According to the ASAS architecture, we grouped the research into several phases. This was done to obtain a clearer picture of how the ASAS system architecture has been developed. The general architecture of the ASAS is mainly divided into five main parts: 1) input, 2) preprocessing, 3) feature extraction, 4) scoring, and 5) output generation. We will discuss these parts in the following sections. We illustrate the general architecture of the ASAS system in Fig. 4.

J. of Soft Computing and Data Mining Vol. 6 No. 3 (2025) p. 323-355

Fig. 4 *General architecture of automatic short answer scoring (ASAS)*

5.1 Data Collection and Preprocessing Phase

The ASAS process starts by collecting student responses. Often, manually graded answers serve as training data. Researchers use different preprocessing steps to prepare the questions, student answers, and reference answers [125]-[134]. The main goal is to improve data quality, turn natural language into a numerical format, and extract features while reducing noise [30, 44, 135]. Common steps include cleaning, normalization, and formatting inputs for the models. Preprocessing is also necessary for other resources, such as handwritten or paper-based answers [136]-[146].

5.1.1 Cleaning and Normalization

One purpose of the preprocessing phase is to ensure data consistency before it is processed in the ASAS model. Several tasks are carried out to obtain clean and consistent question-and-answer data. Case folding or lowercasing involves converting all characters to lowercase, as capitalization is generally not significant in NLP [30, 43, 139]. Removing punctuation and special characters helps to treat each text equally and eliminate meaningless characters [44, 98, 136]. Furthermore, stop word removal eliminates common words like “the”, “a” or “is” that

contribute little to the meaning of a question and answer and appear frequently across the text [72, 84]. Spelling correction and symbol and abbreviation normalization are also applied to improve data consistency [147].

5.1.2 Tokenization and Segmentation

The tokenization phase splits the question and answer text into smaller units, like words or subwords [30, 135]. Several ASAS systems also performed sentence tokenization to divide the question-and-answer text into individual sentences [148]-[154]. Languages without whitespace separators, like Chinese, required specific word segmentation tools [70, 155].

5.1.3 Lexical and Morphological Normalization

Question-and-answer text must be consistent in both vocabulary and word structure [156]-[176]. Several NLP techniques, such as stemming, help to standardize different inflections of the same word by reducing words to their root form. Specific algorithms, such as Porter Stemming and Nazief-Adriani, were used to remove prefixes and suffixes [29, 62, 136]. Lemmatization was also used to transform question and answer words into their meaningful base form, known as the lemma [136, 177]. An abbreviation checker can also be applied to further clean answers [98].

5.1.4 Input Formatting for Model

The important part of the question and answer text should be easily identified by the model. Concatenation of student answers, reference answers, and sometimes questions, into a single input sequence is common for deep learning models [7, 37, 170]. Special tokens like "[CLS]" and "[SEP]" are utilized with transformer models to indicate the beginning and end of sequences, as well as to separate different components of the input [5, 14, 52, 79]. Padding and truncation mechanisms were also applied to ensure that input sequences have a fixed length, which is required for many neural network architectures [92].

5.1.5 Question and Answer Filtering

Questions, reference answers, and student answers text should be filtered to ensure a fair grading process. Deleting words or phrases in a student's answers that are directly connected to the solution or question can be applied as a custom filtering rule [177]-[183]. Question demoting mechanisms were also used to exclude words already present in the question, thus preventing models from unfairly rewarding students who repeat parts of the question [44]. For programming-related answers, models may need to handle non-textual characters like symbols [21, 23]. Sentence simplification methods were also applied using LLMs to reduce the answer's length while still preserving the content [73]. Another filtering method was removing zero-value answers, which indicates that the student did not answer the question [93].

5.1.6 Preprocessing Other Resources

Other resources (e.g., External Corpora, Lecture Notes) need to be handled. For paper-based tests or handwritten responses, optical character recognition (OCR) is used to convert images of text into a digital format, followed by structuring the words into paragraphs and lines [36, 146]. External corpora like Wikipedia, used for pre-training models, undergo similar steps such as lowercasing, punctuation and number removal, stop word removal, and stemming/lemmatization to clean and prepare the data [136].

While ASAS systems were collecting reference answers from community question answering websites like Quora or Stack Overflow, keywords were extracted and weighted based on their parts of speech (PoS) to select relevant content [5].

5.2 Feature Extraction

The next stage of ASAS is to extract important features from the question and answer that models use to analyze and capture context and meaning. Below are the feature extraction approaches developed by researchers.

- **Keyword Extraction.** The keyword extraction mechanism identifies important terms from answer and question texts using algorithms like Rapid Automatic Keyword Extraction (RAKE) [68, 84, 127, 144, 168].
- **Word Embeddings/Vectorization.** The embedding mechanism transforms question and answer texts into numerical vectors, capturing the semantic meaning and relationships between words or phrases. This includes Word2Vec [4, 150], TF-IDF [30], BERT [52], Glove [59], and GPT [59]).
- **Part-of-Speech (PoS) Tagging.** PoS tagging mechanism identifies the grammatical role of each word within a question or answer sentence [30, 110, 136].

- **Named Entity Recognition (NER).** This mechanism identifies and classifies named entities (e.g., persons, organizations, locations) within the text [136, 164].
- **Dependency Parsing.** This mechanism scrutinizes grammatical relationships within question or answer sentences [136].
- **N-gram.** This mechanism sequences N-consecutive words or characters as a feature. And it is also used to capture local word order and frequency patterns [164].
- **Topic Modelling**. This mechanism is used to uncover underlying themes or concepts in student and ideal answers, which are then compared for thematic similarity [30, 112].
- **Question Answering (QA) Models.** Question answering models, such as T5, are applied to extract answers and ideal answers for further similarity evaluation [30].
- **Semantic Facet Extraction.** This mechanism is used to identify fine-grained semantic components relevant to learning objectives [117].

Researchers have extracted a range of features from the question-and-answer text using various approaches. The most commonly used features are keywords and word embeddings. Features such as NER, topic modelling, word dependency, and PoS play a supporting role. These features will be used in the next phase, score generation. However, challenges in feature extraction methods, such as semantic facet extraction, remain ripe for further exploration.

5.3 Scoring Phase

Early approaches captured vocabulary, sentence, and chapter-level features using CNN and BiLSTM networks to enhance scoring accuracy [166]. Meanwhile, deep neural networks required more data to gain optimal performance [128]. The performance of automated text-based evaluation systems was generally measured using metrics such as Quadratic Weighted Kappa (QWK) [72, 82, 91]. Next, the ASAS architecture in terms of how it performs the scoring phase. A research [83] mentioned that there are two types of scores: the first is "binary scoring", which focuses on conceptual questions with answers that can be graded as either correct or incorrect, thus facilitating a binary classification task.; and the second is "complex activities", the quiz that involves more complex activities with varying student participation and is scored on a scale of 0 to 5 or 0 to 6 points. Meanwhile, research [106] scored holistic rubrics with score points such as 0, 1, 2, etc. However, to produce a score, the ASAS model requires guidelines for assessment methods. Generally, the scoring guideline method can be classified into two main categories: scoring with references and scoring without references. We illustrate the scoring method categories in Fig. 5.

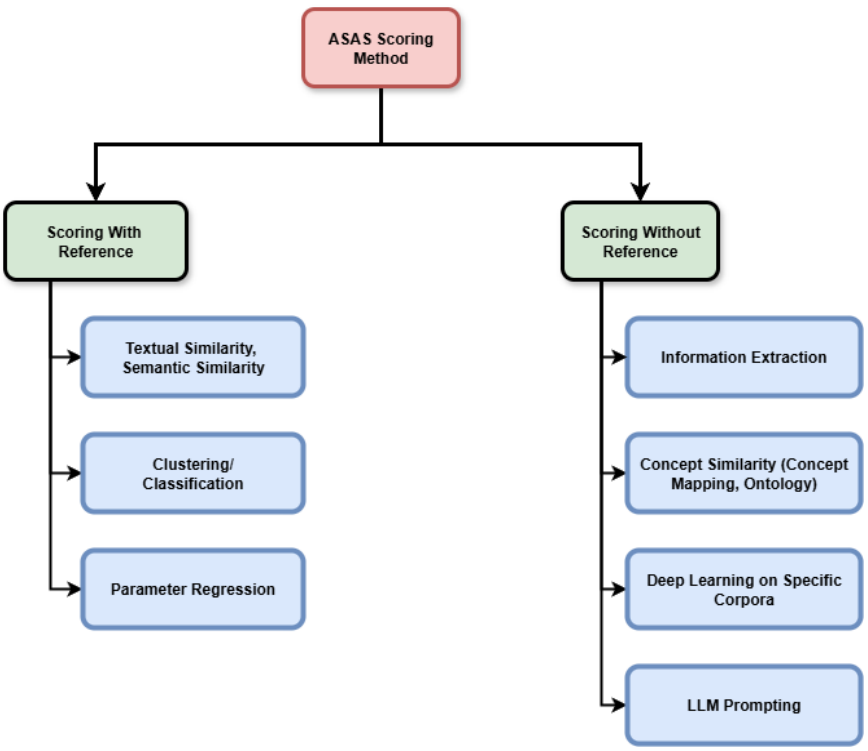

Fig. 5 *ASAS scoring method categories*

5.3.1 Scoring With References

Scoring with references is a condition where a reference answer exists. These reference answers can be correct, incorrect, partially correct, or partially incorrect, or answers with a specific score. This answer is usually obtained from human graders, experts, or teachers. The scoring method for this reference uses text feature extraction, with the results used to compute the score. From the studies conducted, it is noteworthy that the scoring method with references is carried out in the following ways:

- ***Scoring with Similarity.*** Similarity is scored by comparing students' answers with the golden answers. This is done by calculating similarity based on sentence statistics using TF-IDF or semantic similarity using word embedding, sentence embedding, or sense embedding methods. Deep learning models play a role in providing semantic weights and are trained to identify similarities based on semantic weights [4, 35, 116, 132, 147, 164, 171].
- ***Scoring with Clustering/Classification.*** Scoring with clustering or classification is performed by first determining the desired clusters of scores. The researcher used several types of cluster scores, such as the 2-way score (for example, true or false), 3-way score (for example, wrong, partially correct, and correct), or 5-way score (for example, 1 to 5 points) [185]. Deep learning models play a role in forming clusters using clustering algorithms such as K-Nearest Neighbors [16, 33, 113, 168]. A research [165] introduced a novel metric learning (MeL)-based pre-training method for answer representation optimization. This strategy encourages the clustering of similar representations while pushing dissimilar ones apart, thereby facilitating the formation of a more coherent same-score and distinct different-score answer embedding space.
- ***Scoring with Regression.*** Scoring with regression is used when many considerations are involved in the score, such as the proximity of the answer's meaning, the presence of certain keywords, or the number of words. These parameters can be calculated using a regression approach. In this case, the deep learning model identifies the score-determining points and then calculates the regression using algorithms such as linear regression [114, 148]. Commonly used regression models include Support Vector Regression (SVR) [43, 73], Linear Regression (LR) [44, 73], Ridge Regression (RR) [44, 73], Random Forest Regression (RF)[43], ElasticNet (EN) [43], Isotonic Regression (IR) [73], AdaBoost Regressor (Ada) [43], and a Deep Learning based model like LSTM [151] and BERT [78, 136] that is configured as a regressor. Nampoothiri et al. [100] explore automating short-answer grading using machine learning and NLP. It uses transformer models (BERT, SBERT, T5, Genism, XLNet) to convert text into numerical representations and regression techniques (Linear Regression, SVM, Random Forest, AdaBoost) to estimate scores. The BERT model combined with Random Forest performs best, showing superior accuracy. The primary challenge is the need for an additional determinator to map real numbers to integer values, as human graders typically assign integer scores [151].

5.3.2 Scoring Without References

Scoring without references is carried out using concept similarity techniques, such as concept mapping [28], ontology [127], or deep learning of a model on specific corpora. Another way is to prompt either zero-shot or few-shot learning from a particular LLM. Researchers [22] introduced an ASAS framework aimed at learning from unlabelled domain-specific contexts and minimizing both training and inference times. Their framework consists of three main phases. In the first phase, the framework constructs an unlabelled domain-specific corpus. In the second phase, the framework trains a pre-trained language model using the corpus generated in the first phase. And the third phase, the framework fine-tunes the pre-trained model with a limited amount of labelled domain-specific data [22]. They explored various domain adaptations using two types of vocabulary during the pretraining. ASAS systems increasingly employ unsupervised and generative approaches to score responses without requiring a gold-standard reference [22]. This allows grading of open-ended and divergent responses when fixed-reference answers are impractical.

Transformer-based models (e.g., GPT and BERT) are used to analyse the semantic meaning of answers rather than matching against a reference. These models focus on the coherence, relevance, and logical flow of the responses. This allows grading of creative, argumentative, or opinion-based questions in which multiple correct interpretations may exist. To assess answer quality, scoring without reference approaches often employs feature extraction methods, such as syntactic complexity, semantic similarity, sentiment analysis, and topic modelling. Tools like WordNet were not yet available for some languages, such as Indonesian and Arabic, which limits their use in semantic network-based preprocessing. The scarcity of linguistic resources in Arabic, including corpora, dictionaries, and lexicons, presents significant challenges for ASAS research, as transformers require significant data to train effectively [68, 109].

Other researchers [142] utilized LLMs to score rubric-based assessments and provide feedback for students. LLMs were prompted to act as expert examiners. They result in a strong correlation between the scores generated by LLMs and those by human examiners. Their research showed potential future advancements of LLMs in ASAS

systems. They also acknowledged LLMs' limitations related to the need for clear rubrics. Their research also concludes that AI-assisted scoring tools can lessen grading workloads and improve writing assessment in education.

5.3.3 Output and Post-Processing Phase

The ASAS system generally results in a predicted score for the student's answer. More advanced systems may yield more advanced outputs, such as identifying discrepancies between automated and human scores.
The other advanced output is feedback generation, which provides comprehensive feedback to teachers or students to improve the learning experience. Adaptive learning integrations were also among the advanced outputs of ASAS, in which ASAS feeds into intelligent tutoring systems to guide students' learning paths. Last, ASAS also outputs prompts to guide the LLM's scoring behaviour with rubrics.

Overall, the architectural advancements in ASAS show a trend towards increasingly context understanding capabilities, moving from explicit feature engineering to implicit feature learning through deep neural networks and, most recently, leveraging the pre-trained knowledge and in-context learning abilities of large language models. The quality and quantity of training data are critical. Thus, inadequate or biased training data for ASAS can lead to poor evaluation results [30]. ASAS datasets are often unbalanced in terms of the quantity of different grade clusters, which can negatively affect model performance [23]. Some question and answer datasets also contain more unfamiliar words, which can impact the effectiveness of certain preprocessing methods [136]. Models can also struggle with misclassified answer data containing multiple invalid words or phrases [149].

6. Technological Approach of ASAS (RQ2)

In this study, we identified six distinct technology clusters used for ASAS. Each cluster is distinguished by its unique language features and scoring techniques. A detailed illustration of these ASAS technology categorizations is shown in Fig. 6.

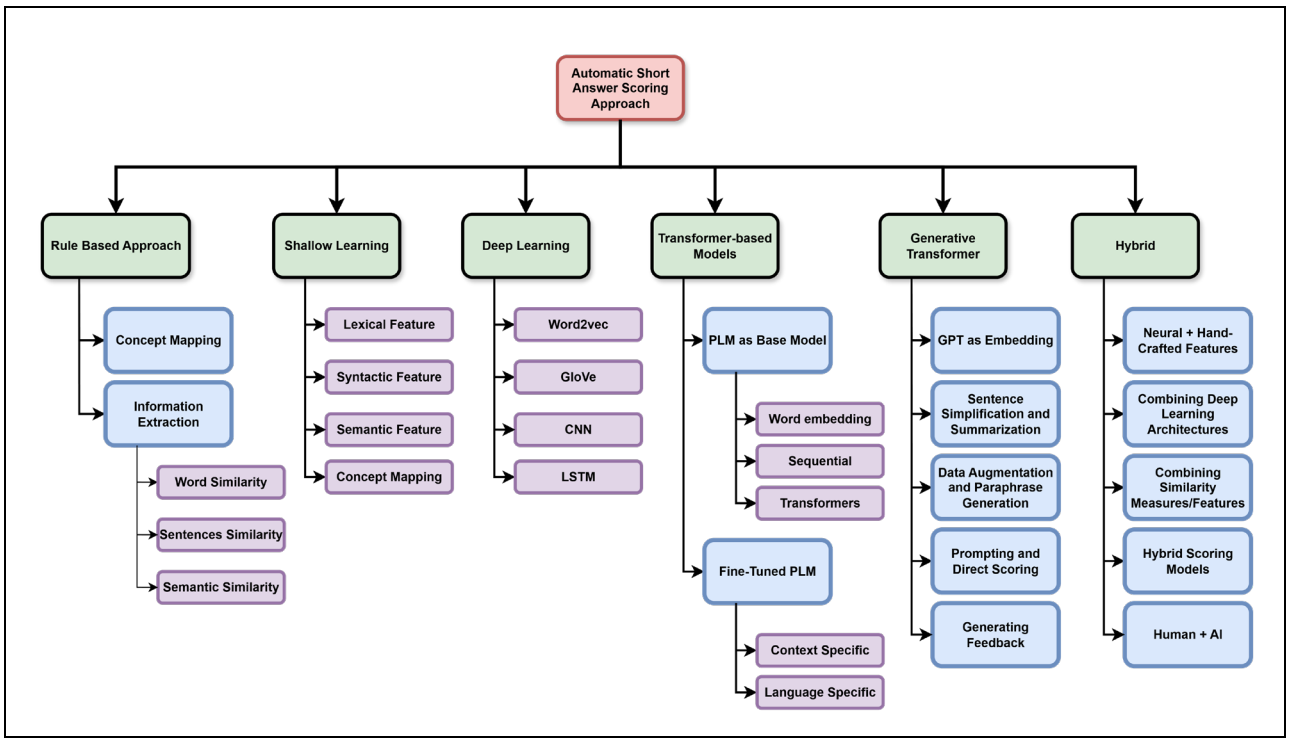

Fig. 6 *Technological approaches in short answer scoring*

6.1 Rule-Based Approach

The rule-based approach was also categorized as a traditional approach, which prioritizes feature extraction from text [15]. Extracted text features were used to score the short-answer responses using several methods, such as statistical calculation, semantic similarity, conceptual similarity, or simple answer clustering [65, 105]. Features that were used for scoring calculation include keyword statistics, keyword similarity, parts-of-speech, and Named Entity Recognition (NER) [178]. Similarity-based approaches have also emerged to measure the distance between a student's response and reference answers, using techniques such as Latent Semantic Analysis (LSA) [39].

While the prevalent method for evaluating reference-based answer scoring models involves assessing the degree of similarity between the student's response and the reference answer, as proposed by [21], others [182] used a Semantic Feature-Wise approach to map the relationship of the question, the student's answer, and the teacher's reference answers. Rule-based approaches are early systems that rely on pattern matching and semantic similarity, but are limited in their ability to capture diverse expressions.

6.2 Machine Learning Approach

Researchers use Machine Learning (ML) approaches to improve the accuracy of ASAS. ML is used by training the models on the prepared data. The main distinction between the ML and DL approaches is that the ML approach uses textual features to directly compute the score, whereas the DL approach employs a more complex, deeper contextual understanding to compute the score.

Common machine learning algorithms employed include SVM, Random Forest, Linear Regression, and KNN [11, 43, 44]. Machine learning approaches have been used in [86] to actively train the model to identify correct and incorrect answers. Other research [3] uses Q-Learning and synset-based language processing to enhance grading performance incrementally. A method using "SemSpace Sense Vector" and MaLSTM was proposed in [160], which was a vector based on concepts from the WordNet lexical-semantic network.

6.3 Deep Learning Approach

A major improvement was achieved with the Deep Learning (DL) approach. In the DL approach, models learn directly from raw data without manual feature engineering. DL uses word embeddings to represent question and answer words in a vector space, capturing their semantic or contextual use. These embeddings can be pre-trained on large, unlabeled datasets, such as Word2vec and GloVe, or fine-tuned by models like ELMo. Models that are based on CNN and LSTM networks, including Bi-LSTM versions, are used to capture local, global, and sequential context information in question and answer text [72, 146, 163].

A supervised machine learning method for ASAS was presented in [153]. They compared CNN, Char CNN, LSTM, and BERT models. Their research results show that BERT outperformed CNN and LSTM. This is because BERT's deep contextual representations of question and answer consider context from both directions, which allows the context to be understood better [153].

An ASAS system that uses the XLNet model was introduced in [124] to turn students' answers into vectors that capture context and meaning. These vectors train an LSTM model to calculate sentence coherence. The system performed better, achieving a high QWK score and surpassing current baseline models.

6.4 Transformer-based Models Approach

Recently, researchers have increasingly used Transformer-based models to encode answer text dynamically. These models address small-dataset issues by leveraging pre-trained knowledge and fine-tuning on a specific domain. This process often involves complex networks like CNNs and LSTMs [181]. Such models can be combined or configured in Siamese architectures to encode sequences and capture semantic details. Some also use attention mechanisms highlight important text parts. Most Pre-trained Language Models (PLMs) use neural networks tailored to the embedding type. Word embeddings map words to numbers in a semantic space, capturing meaning and context.

Although the quality of these embeddings generally improves with larger training datasets, other factors also influence their suitability for specific tasks. The characteristics of the training data determine the nature of an embedding, because the meaning of a word can change depending on its context [106].

PLMs have been widely used. Most researchers use PLMs, such as BERT, as a model to tokenize questions, reference answers, and student answers. The results of tokenization are then used in subsequent processes, such as calculating the score using cosine similarity or clustering the answer as true, partially true, or false. Fine-tuning in PLM was also performed to improve the accuracy of the ASAS.

Pre-trained models such as RoBERTa, DistilBERT, T5, and XLNet are commonly used as base models and then fine-tuned to achieve the highest accuracy. Some studies have used domain-specific datasets for fine-tuning. The use of these domain-specific datasets has been shown to outperform base models derived from large datasets. The results of the study also showed different effects across languages.

Many languages have been used in research models. We have identified some studies that explore the use of the ASAS in specific languages, which we will explain in the next section. We found that PLM can be utilized in various ways. Some researchers use PLM as a base model, while others conduct fine-tuning to achieve greater accuracy.

6.4.1 PLMs as a base model

Researchers widely use the PLMs to tokenize and embed questions, reference answers, and student answers. Embedding is conducted on question or answer words and sentences. Researchers have focused on the external factors of PLMs that can affect ASAS systems, such as meta-learning [52, 167]; curriculum learning [180]; and scoring rubrics [11, 41, 68].

A strategy that matches the features of answers with the criteria stated in a rubric was introduced in [11]. The shortcut learning behaviour was successfully reduced, and the model's trustworthiness was improved through their approach. The proposed strategy [11] includes a detection mechanism that uses a feature attribution approach to identify superficial behaviour. Additionally, a correction method [11] was also implemented to retrain the model, ensuring it aligns with the rubric-related annotations and mitigating the impact of these superficial cues.

6.4.2 Fine-tuned PLM

Researchers have performed many fine-tuning methods in PLM to provide a contextual understanding of the model. The created dataset varied according to the theme developed. They collected questions on a specific domain, obtained the ideal answer from experts, and then trained the model based on that data. Some researchers have also used transformer models to generate paraphrases from ideal answers to improve model training accuracy. This study highlights several effective strategies for fine-tuning the ASAS, which can enhance its performance, including data augmentation techniques such as synonym replacement [21]; Back-translation and paraphrasing to expand training datasets [109]; Simplification and normalization, preprocessing through token normalization, spelling correction, and handling noise; contrastive learning, models are trained to distinguish between correct and incorrect answers [99].

6.5 Generative Pre-Trained Transformer (GPT) Approach

Generative Pre-Trained/LLM-based approaches leverage LLMs directly for text embedding or for a completion-based ASAS system by providing the model with examples of questions, answers, and scores in the prompt itself. These models can bypass traditional training for new tasks through in-context learning. On top of that, some researchers have used GPT to generate embeddings before scoring. Other researchers have used GPT to generate scores directly from students' answers.

6.5.1 GPT as an Embedding Generator

Early transformer models like ELMo, BERT, GPT, and GPT-2 have been used for pre-training the embedding in ASAS's tasks [5, 141]. Specific GPT variants like Sentence-GPT (SGPT) were used to encode sentences for semantic similarity tasks in ASAS [26]. The text-embedding-ada-002 (EADA2) model, based on generative AI, is also used to create contextual similarity embeddings between student answers and model keys [73].

GPT models were also used in [156] to increase scoring accuracy and alignment with human raters. Their approach shows the ability to capture contextual relations, thereby enhancing the content and quality understanding of an essay. Their methods incorporate linguistic features such as concreteness, uncertainty, conviction, commitment, and writing styles into transformer-based models.

6.5.2 GPT as Sentence Simplification and Summarization

LLMs such as GPT-3.5 were integrated into ASAS systems in [73] to shorten or simplify student answers while retaining the meaning. These simplification methods helped to examine poorly structured or wordy student answers.

6.5.3 Data Augmentation and Paraphrase Generation

A novel ASAS approach was presented in [45, 118] by leveraging LLMs as an augmentation tool. Retrieval Augmentation Generation (RAG) techniques were implemented by providing LLMs with access to an external knowledge database and using it to generate data. They proposed this approach to enhance ASAS models with greater consistency and reliability, and to reduce hallucinations. The external knowledge databases include the short-answer dataset (a sample of questions, answers, and scores), lecture notes or course materials, and a context-specific dataset. With a more appropriate context provided by the retrieved texts, LLMs generate higher-quality data. An example of the generated prompt is shown in Fig.7 [45].

> ***General instruction:*** *You are the instructor of a college-level Introductory Biology course. You are going to grade the exam for this course. Your grading should be based on the question asked, the full-credit answer, the student's answer, and nothing else. Give the binary score 1 or 0, in which 1 means the student's answer is correct and 0 means the student's answer is incorrect or does not answer the question, and justify your grading.*
>
> ***Question-specific instruction:*** *As long as the answer mentions or implies that the molecule contains just carbon, it should be considered as being correct and graded as 1.*

Fig. 7 *Sample of scoring with a prompt* [45]

Researchers have also used RAG to generate paraphrases of questions or answers [73, 109]. The purpose of paraphrasing is to reduce bias in the question or the student's answer. Paraphrases are also used to provide more diverse reference answers to improve accuracy. Objects used for paraphrasing generation include question and/or answer variation [109] and dataset modification [169]. The results of the RAG for paraphrase generation were used as inputs to the model training and were also used to generate the score, or were directly used in the similarity or classification phase to calculate the score. We illustrate the differences in various RAG usages for prompt and paraphrase generation in Fig. 8. Fig. 8(a) shows RAG for prompt-scoring generation, Fig. 8(b) shows RAG for reference-answer modification, and Fig. 8(c) shows RAG for student-answer modification.

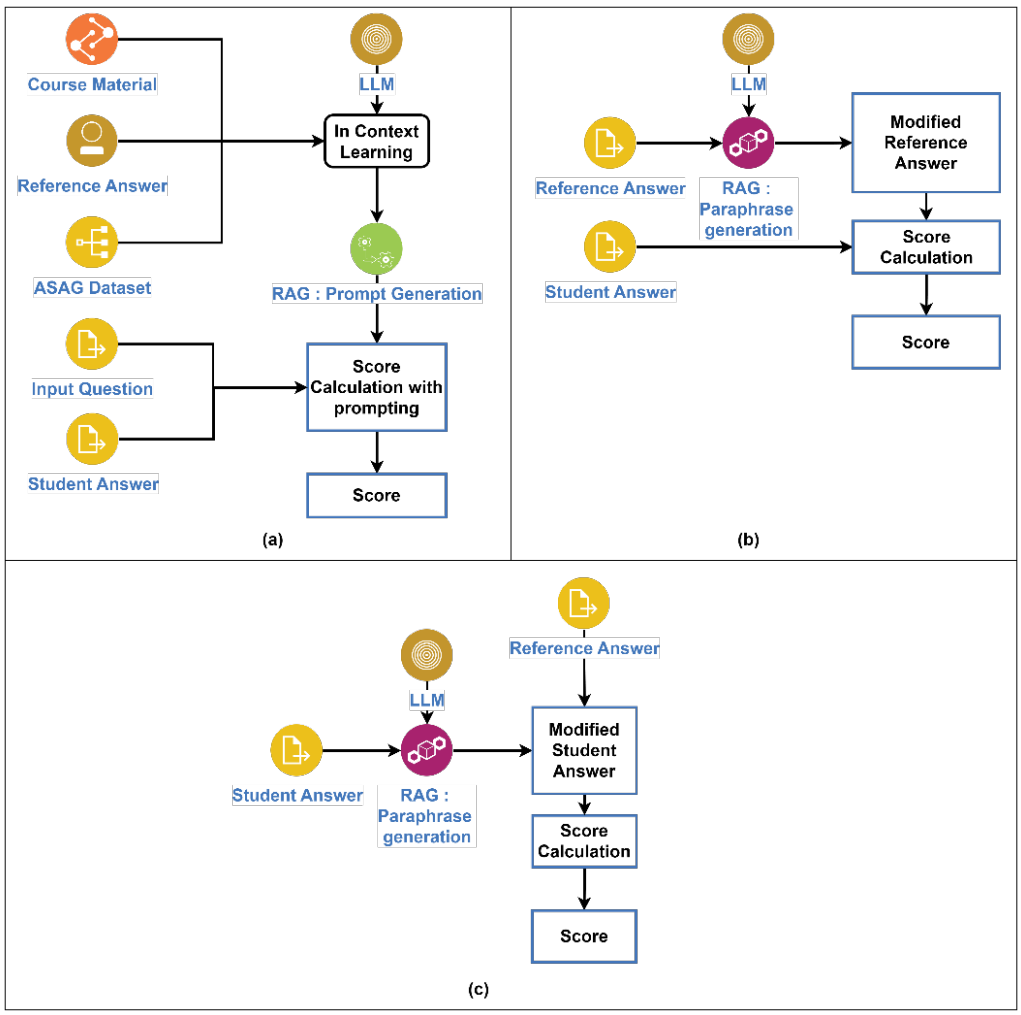

Fig. 8 *The use of retrieval augmentation generation (RAG)*

6.5.4 GPT for Prompting and Direct Scoring

The study of [40] used chain-of-thought prompting, which provides a reasoning chain along with an answer to the labelled instances used for in-context learning. This method enables the model to generate accurate inferences, incorporating reasoning alongside the answers it provides. It has been noted that chain-of-thought reasoning is especially beneficial for formative assessment scoring in science. This is because the open-ended nature of the questions can complicate alignment in scoring, even among human evaluators.

Several studies [9, 52, 53, 65, 69, 75] evaluated GPT-3.5 and GPT-4 in zero-shot and few-shot settings for ASAS tasks across various languages and subjects, including English, German, and Finnish.

While some early findings suggested GPT could not be used directly, on the contrary, later research showed its potential, especially with improved prompt components [53, 75].

GPT-4 can function as an effective short-answer grader. It achieves significant agreement with human graders, particularly when provided with the right problem formulation and scoring examples (few-shot learning). It can also follow scoring rubrics and identify correct or missing information in student responses [53, 75].

6.5.5 GPT for Generating Feedback

Researchers used GPT models, particularly GPT-3.5 and GPT-4, to generate detailed feedback on inconsistent answers and provide tailored feedback to students. This generated feedback can assist novice graders and boost the performance and consistency of expert graders [52, 65, 118, 151].

In conclusion, Transformer-based models have been widely used and have improved the ASAS system. They can capture complex language details and meaning. These models have become a common approach in ASAS and outperform traditional machine learning methods. While transformer-based models can be trained directly for ASAS, the best performances were achieved when combining their semantic descriptions with carefully hand-engineered features. However, the study indicated that GPT-4's performance is comparable to that of hand-engineered models but falls short when compared to pre-trained LLMs that undergo specialized training [89]. Transformer-based models used for ASAS are described in Table 4.

Table 4 *Transformer-based models used for ASAS*

Transformer Models	References
BERT	[78, 97, 101, 143, 115, 152, 154, 157, 170, 173, 185]
RoBERTa	[80]
ALBERT	[21]
Hybrid BERT with Neural Networks	[72, 79, 185]
XLNet	[62, 149]
MPNet	[78, 126]
ELMo	[65]
SBERT	[21, 41, 111]
T5 (Text-to-Text Transformers)	[30]
GPT 3.5	[45, 65]
GPT 4	[45, 65, 75, 89]
ChatGPT	[31, 74]

6.6 Hybrid Approach

Some researchers have combined GPT and PLM or traditional approaches by proposing several types of frameworks. As mentioned in [169], GPT was used to generate paraphrase reference answers and student answers to balance the dataset before it was used for training the PLM. An encoder-decoder model named ARAG-ED was introduced in [109], which was designed using GPT to generate paraphrased reference answers based on the initial reference answer. They then used the PLM to perform the scoring task. Another study [59] used GPT for sentence embedding, and then the results were used to generate scores via regression, such as Cosine Similarity. A hybrid approach that combined a CNN Deep Learning layer with BERT and LTSM was proposed in [79] a customized multi-head attention mechanism adapted with BERT was employed. A hybrid deep learning approach that combined BERT and LSTM was proposed in [72] to improve automatic short-answer scoring.

A hybrid scoring model that enhances a pre-trained model based on BERT was proposed in [125]. The model effectively combines answer representations with a vector constructed from carefully crafted essays' linguistic

features, adapted from [161, 162]. The research [125] trained these vectors on an LSTM and a Bi-LSTM model to capture sentence connectivity with other sentence semantics. Hybrid approaches in ASAS combine two or more methods or feature types to enhance performance, accuracy, and robustness. Below are some hybrid approaches researchers have developed to improve ASAS performance.

6.6.1 Combining Neural and Hand-crafted Features

Some researchers successfully combine manually engineered (hand-crafted) features with neural models [12]. Neural models generate dense vector representations of input answers, which can be effectively combined with carefully designed linguistic feature vectors. Studies have demonstrated significant performance increases when hand-crafted features are added to neural representations. This is because knowledge extracted from embedding-based representations can complement previously developed syntactic, semantic, and lexical features. Research [85] found that combining token-level features with embeddings can provide more general, domain-independent information to the model and achieve better results. They used a random-forest feature-based model with a broad set of linguistic features (part-of-speech, keywords with predicted weights, logical operators, and lexical diversity) and pre-trained classical embeddings (word2vec, doc2vec). A hybrid approach for Arabic ASAS was combined manually to extract features, such as word2vec and contextual embedding features using the BERT model, and achieved high accuracy [1].

6.6.2 Combining Different Deep Learning Architectures

Researchers have also proposed hybrid models that combine different deep learning architectures, such as CNNs and LSTMs, to capture different aspects of textual information. Some systems utilize CNN and LSTM models to predict scores, capturing both local and global context information. Some researchers [90, 91] proposed a deep learning approach that combines word embeddings, CNN, and LSTM networks for ASAS. The other researcher [37] The proposed cross-lingual hybrid neural network uses “CNN and bidirectional LSTM (Bi-LSTM) sequentially to encode answer text”. A group of researchers later introduced an interaction enhancement layer with attention mechanisms, and finally used Siamese Bi-LSTM networks to fuse enhanced representations for grade classification [37, 51, 114, 172, 179].

A new hybrid approach combines BERT with a customized multi-head attention layer and a parallel connection of CNN layers to enhance text comprehension and improve performance [37, 79]. This model also integrates BERT, LSTM, and CNN layers. ASAS-F-RAG, a few-shot ASAS-F system, uses LLMs in combination with ColBERT to retrieve the most similar examples from training data to generate grounded feedback. ColBERT encodes input into a matrix of contextual token-level embeddings, capturing fine-grained similarities [52].

A two-phase approach was introduced by [57] to automated short answer scoring that leverages “cross-prompt data” to automated short answer scoring that leverages “cross-prompt data” to reduce the need for extensive in-prompt training data. Their approach involves “pre-finetuning a model on existing rubrics and answers and then finetuning it on a new prompt”, addressing data accessibility issues. BERT and LLMs were used in the experiments [57] on the RIKEN SAS dataset [56], and demonstrated significant improvements in scoring accuracy, particularly in low-resource settings.

A novel workflow that combines active and deep learning was introduced in [107] to improve the accuracy of automated scoring for short-answer questions, while also reducing the cost associated with manual scoring. Results show high accuracy and significant cost reduction compared to traditional methods and other machine learning approaches.

6.6.3 Combining Different Similarity Measures/Features

ASAS systems often combine features in both lexical and semantic forms to provide a final grade. This fusion aims to provide a richer feature matrix for machine learning approaches. Intelligent Descriptive answer E-Assessment System (IDEAS) framework exploits eight particular language features from teachers’ and students’ answers for scoring, integrating the traditional bag-of-words method with a modern sentence embedding approach for semantics [151].

Another proposed hybrid approach utilized word occurrence statistics, TFIDF, Latent Semantic Analysis, semantic similarity using Infersent, and summary similarity as input features for various classification models [110]. Also, in the Arabic ASAS model, there is a study [94] that evaluates text similarity using knowledge-based similarity (e.g., WordNet), and word and contextual embedding similarity (e.g., Word2vec, BERT).

A subjective answer grader that combines scores from model similarity (a linguistic model computing semantic similarity) and keyword-matching similarity was proposed in [84]. The development and evaluation of a short-answer automated grading system for Japanese language learners was proposed by [88]. Using the actual test dataset SIMPLE-O for university students in Indonesia, they combine similarity scoring and binary classification with a BERT model.

The DAES model [30] combines LDA for thematic coverage and T5 for semantic understanding, aggregating their similarity scores with assigned weights. Meanwhile, study [13] presents a new, hybrid method of evaluating short answers in Polish, combining text similarity algorithms with neural networks.

The comparative analysis showed that the hybrid approach outperforms previous methods, offering high precision and recall. Despite the very good results, the authors plan further research, including the use of deep learning methods, to further improve the effectiveness of the assessment.

6.6.4 Hybrid Scoring Models (Classification and Regression)

Some frameworks treat ASAS as a multiclass classification problem where each mark is a distinct label, departing from conventional binary classification or regression approaches. A study [134] addressed scoring as a regression task as well as a classification task; they experimented with "a rich set of text similarity features, CBoW, and TF-IDF to train a machine learning regressor and classifier whose output is then conveyed into a fusion model for the final scoring. The effectiveness of combining BERT with GRU for regression tasks in automated student answer evaluation, particularly on small datasets, was investigated [76]. Results suggest that hybrid models effectively leverage the strengths of both transformer-based and recurrent architectures.

6.6.5 Hybrid Human-AI Systems

ASAS systems are often set up as hybrid systems, providing suggestions that require human review. This approach lets the teacher maintain control and save time by skipping reviews when the predictions are reliable [111, 133, 184]. The ASSIST system proposed by [185] used an interactive response coding method, where raw responses are grouped with expert feedback. The system then learns to assign scores and identify keywords, helping to understand the data.

A Reflective Prompt Engineering (RPE) was introduced in [9] for scoring short answers in biology using Generative Models. RPE uses iterative human-AI collaboration, in which experts guide the LLM by integrating its inferred criteria into subsequent prompts. Hybrid approaches in ASAS represent a continuous effort to combine the strengths of different techniques, ranging from traditional feature engineering to advanced deep learning models and natural language processing methods, to achieve more accurate, reliable, and objective grading outcomes, often aiming for human-like performance [109]. They are also explored to make systems more adaptable, efficient, and applicable across diverse domains and languages.

7. Dataset For Model Training (RQ3)

Researchers have produced various datasets. Both can be used publicly and by researchers themselves to improve ASAS performance. In this study, we categorized the ASAS dataset into a popular public dataset and a new fine-tuned dataset, as shown in Table 5.

Table 5 *Dataset for ASAS*

Dataset	*Year*	*Content Length*	*Scoring type*	*URL*	*References*
Popular Public Dataset					
SciEntsBank	2008	Q: 197 A: 16.000 (aprox.)	2-way, 3-way, 5-way	https://github.com/dbbrandt/short_answer_granding_capstone_project/tree/master/data/source_data/sciEntsBank	[102]
SemEval-2013 dataset	2013	Q: n/a A: 15,357	2-way, 3-way, 5-way	https://www.cs.york.ac.uk/semeval-2013/task4/	[47, 48]
BEETLE dataset	2014	Q: 56 A: 3,000		n/a	[46]
STITA dataset		Q: n/a A: 333	2-way: correct or incorrect	https://github.com/edgresearch/dataset-automaticgrading-2022	[43]
Mohler dataset	2011	Q: 81 A: 2,273	[0-5]	https://aclanthology.org/attachments/P11-1076.Datasets.zip	[96]

ASAP-SAS dataset	2013	A: 17,207 Q: 10	[0-2], or [0-3]	https://www.kaggle.com/competitions/asap-aes	[145]
SQuAD	2016-2018	100.000+ question-answer pairs	Question Answering	https://rajpurkar.github.io/SQuAD-explorer/	[122, 123]
New dataset fine-tuned from a publicly available dataset					
USCIS dataset	2013	Q: 20 A: 698	2-way: correct or incorrect	http://research.microsoft.com/~sumitb/grading	[16]
Indonesian ASAG Datasets	2018	Q: n/a A: 2,162	[0,100]	https://data.mendeley.com/datasets/6gp8m72s9p/1	[120]
RIKEN dataset	2020	Q: n/a A: n/a		https://www.nii.ac.jp/dsc/idr/rdata/RIKEN-SAA/	[56, 95]
CU-NLP, a NLP course of Cukulova University dataset	2021	Q: n/a A: 171	[0-100]	https://bmb.cu.edu.tr/uorhan/CuNLP.htm	[160]
AR-ASAG, an Arabic language dataset	2021	Q: 48 A: 2,133	[0-5]	https://data.mendeley.com/datasets/dj95jh332j/1	[108, 119]
AraScore-Dataset, an Arabic language dataset	2021	Q: 5 A: 2,500	[0-3]	https://github.com/guc-research/AraScore-Dataset	[50]
Bilingual Short Answer Feedback (SAF) dataset	2022	Q: 22 A: 4,519	[0-1]	https://github.com/SebOchs/SAF	[54]
Short Programming-Related Answer Grading (SPRAG)	2024	Q: 114 A: 4,039	2-way: correct or incorrect	https://github.com/sridevibonthu/SPRAG/tree/main/Data	[23]
IDEAS-ASAG-DATA	2024	Q: 20 A: 800	[0-4]	https://github.com/dbbrandt/short_answer_granding_capstone_project/tree/master/data/seb	[151]

7.1 Context-Specific Fine-Tuning the Transformer-Based Model for ASAS

When building a context-specific ASAS model, the dataset's size during training can limit the model's performance. In contrast, humans can learn to grade effectively from a limited number of labeled student responses. This ability arises from their extensive domain-general knowledge, which enables them to quickly comprehend and adapt the grading criteria to the specific question at hand. A study [181] found that "both domain-general and domain-specific information are integrated during the feature engineering process".

Research [23] has indicated that the development of domain-specific pre-trained language models (PLMs) significantly optimizes both time and resources for developers and researchers engaged in natural language processing tasks within a particular domain. By utilizing domain-specific PLMs, researchers and developers can effectively fine-tune them on smaller datasets, thereby eliminating the need to train a language model from scratch.

The ColBERT model was introduced [52], which efficiently retrieves context-relevant examples from the training data. Unlike traditional models, which encode the entire input into a single vector, ColBERT encodes the input into a matrix of contextual token-level embeddings.

Notably, domain-specific information and contextual embedding techniques can lead to faster ASAS model development and improved performance on domain-specific ASAS tasks. Various ASAS context-specific models have been developed by researchers, which are described in Table 6.

Table 6 *Specific context for ASAS*

Context/Course	Research
Government and Policy	[152]
Software Engineering Course	[23, 45]
Computer Science Course	[137]
Information Technology Course	[98]
Data Mining Course	[159]
Medical Education Course	[39, 65, 142]
Health Informatics Course	[42]
Psychology	[152]
Electronics Course	[25, 77]
Electrical Course	[77]
Mechanical Course	[77]
Science Course	[135, 180]
English Course	[75, 180]
Biology Course	[75, 140]
Introductory Physics Course	[176]
E-Business Course	[164]

However, while context-specific datasets outperform large corpora in performance, they still exhibit limited sentence-level variation. According study [98], improving datasets by including a wider range of question types can better match the different answers students provided for short-answer questions. Also, employing query expansion techniques, such as synonyms and polysemous words, can help find sentences that express the same meaning but are structured differently. This approach has the potential to improve overall performance.

7.2 Language-specific Fine-Tuning Transformer-Based Model for ASAS

A range of specialized languages has been created for ASAS. These languages use advanced transformer-based language models to evaluate and score short answers. They improve the efficiency and objectivity of the assessment process. Researchers have also investigated whether specific languages can affect the accuracy of the ASAS system. The following studies train models in specific languages and multilingual capabilities.

- **English**: The use of pre-trained transformer models such as BERT, RoBERTa, and DistilBERT is common in ASAS research, especially in English. Many of these models are pre-trained on large corpora such as Wikipedia, which mainly features English. As a result, researchers often fine-tune these models [169], suggest new architectures [144], create frameworks [43, 58, 87, 144, 159], and develop domain-specific datasets for their studies, such as programming courses [23] based on English.
- **Arabic**: Efforts have been made to adapt the ASAS systems for Arabic. This includes using Machine Learning [1] and Deep Learning methods to achieve results comparable to those in English [135, 150]. These efforts also involve generating datasets that are not widely available in Arabic [50, 109].
- **Japanese**: ASAS systems have been created for Japanese learners using different approaches like TF-IDF, latent semantic analysis [116], K-means clustering [130], and CNN-LSTM [104]. Researchers also developed advancements in ASAS frameworks and directly implemented them in universities [71, 103]. They have also developed Japanese language datasets to assist native learners with ASAS tasks [56, 95].
- **Indonesian:** The ASAS system in the Bahasa Indonesia Language has evolved from a statistical approach [129], semantic analysis [68], clustering techniques [131], machine learning [10], deep learning [148], language models [170], and hybrid methods. The hybrid methods were developed by combining fine-tuned Indonesian language models, statistical features, machine learning, and deep learning techniques

[67]. Researchers have also examined aspects that could improve the accuracy of the ASAS model in Bahasa Indonesia, including abbreviations, word synonyms [29], transfer learning [172], rubrics [68], and specific material courses [136].

- **Other Languages**: Due to limited research on the following languages, we group other languages in this study, which include Chinese [174], German [111], Portuguese [53, 147], Greek [97], Moroccan [2], Spanish [138], Finnish [32, 32], Dutch [19], Indian[137, 143], Hebrew [140], Polish [13], and Turkish[38].
- **Multilingual Capabilities**: A cross-lingual ASAS system was investigated in [70]which involves scoring models trained in one language on data in another language. The research [70] studied data from five languages, including Chinese, English, French, German, and Spanish. They used both machine translation and multilingual transformer models for their analysis.

However, various studies have shown that transformer-based translation performs worse than models specifically trained for understanding certain languages. Thus, continuous models training on evolving datasets could enhance adaptability to changing linguistic patterns [4].

8. Recent Framework of the ASAS System (RQ4)

Recent ASAS systems are increasingly using a modular design to separate important steps, such as data preprocessing, feature extraction, model inference, and result interpretation. This design improves scalability, allows upgrades to components, such as replacing traditional models with transformer-based models, and helps with better error localization. This modular approach then triggers the development of several ASAS frameworks that are outlined in Table 7.

Table 7 *ASAS framework and architecture*

Name of ASAS Framework	Reference	Year
ESAS	[175]	2020
SmartScore	[158]	2020
SFRN	[182]	2021
Indonesian ASAG system	[136]	2022
iGrade	[6]	2022
SAGAL	[86]	2022
ASAGer	[73]	2023
Ex-ASAG	[159]	2023
Obmaaq	[127]	2023
ASSIST	[92]	2024
DAES (Descriptive Answer Evaluation System)	[30]	2024
IntelliGrader (IDEAS)	[151]	2024
HTL-ASAS	[93]	2024
SteLLA	[118]	2024
EduGuard	[121]	2025

A new framework called “Short Answer Grading with Active Learning (SAGAL)” was introduced in [86]. This framework provides an approach that includes rules for identifying examples of good, borderline, and unusual answers. Based on active learning principles, SAGAL repeatedly selects high-quality examples and requests annotations to improve sampling accuracy [86].

A framework called the “Intelligent Descriptive Answer E-Assessment System (IDEAS)” was introduced [151]. They employed a model-answer-based method that applies eight similarity metrics to compare student answers against model answers. These metrics were derived from a combination of statistical methods and deep learning techniques [151].

A new framework named “Hybrid Transfer Learning for Automated Short-Answer Scoring (HTL-ASAS)” was introduced in [93], which combined different tokenizers from pre-trained models. They also created a new dataset of student responses in the “Introductory Information Technology” course, which teachers manually graded.

A framework called the “EduGuard” was introduced in [121]. This framework is an AI-based online exam management system that offers a secure, automated, and efficient solution for ASAS tasks. EduGuard addressed

issues such as cheating and grading errors through AI-powered proctoring, automated grading, and robust security measures, and delivered real-time monitoring, encryption, and role-based access control.

Modern ASAS frameworks continuously use transformer-based models, as these models can capture context and semantic meaning from text. However, challenges remain in developing transformer-based ASAS frameworks that combine the best elements for optimal ASAS performance. The use of LLMs and GPTs within the ASAS framework also presents additional opportunities for ASAS research and development.

9. Discussion

This review highlights the technological evolution of ASAS systems. We observed that manual grading has become impractical due to time constraints and susceptibility to human bias. Transformer-based models often act as a similarity calculator. LLMs have great potential for providing score feedback, but still struggle to match the accuracy of transformer-based models. LLMs also need to gain a better understanding of context to reduce bias and produce a reasonable score. Meanwhile, current AI technologies offer advanced capabilities, like reasoning and augmentation. Consequently, there is a growing consensus that integrating LLMs and AI is essential for ASAS's future. Although there is a number of emerging research on ASAS technological advancements, there are still areas worth further exploration. These include:

1. **Domain-Specific Constraints**: ASAS systems perform well while trained on a large domain of knowledge, but fail when applied to a specific context. Future research should contribute to methods that improve cross-domain context understanding.
2. **Data Limitations**: Unbalanced and scarce data remain a major challenge for ASAS models, since ASAS have a limited textual resource. Consequently, there is a need for more studies on effective data augmentation techniques.
3. **Large Language Models**: Models like GPT-4 show promise in a better learning experience through scoring feedback, but the models still require balanced fine-tuning and better reasoning. Research into models' performance on vast, diverse corpora and various reasoning methods is still necessary.
4. **Hybrid Models**: Combining different machine learning methods, such as hybrid transfer learning with GPTs, could boost system performance. This combination may create a more robust and trustworthy ASAS system.

Despite advancements in ASAS technologies, our review also uncovers significant gaps that remain and require further attention. These gaps include the need for further studies on how ASAS could provide a fairer and more reasonable score. Additionally, there is a need for further research on how ASAS affects students' learning experiences with personalized feedback. As ASAS's technologies improve, it is also important to examine the ethical issues related to the use of the ASAS system in educational settings.

10. Conclusion

This review clearly maps the current state and evolution of the ASAS system. We observed a shift from traditional, feature-based methods to advanced transformer-based deep learning. Thus, we manage to answer our research questions: (i) Modern ASAS architectures tend to be modular. This modular approach helps researchers to easily arrange the best combination for the ASAS framework. The framework designed by researchers performs better when they use deep learning for feature extraction rather than manual engineering. Therefore, there are potential future studies in combining deep learning with LLMs **(RQ1)**. (ii) Transformer-based models are crucial for ASAS, as they effectively create contextual understanding through sentence embeddings. Recently, LLMs like GPT-4 have been used to provide score feedback, despite generating scores. However, transformer-based models generally serve only as similarity calculators. Thus, there is a need for further research in utilizing transformer-based models to generate more fair and reasonable scores **(RQ2)**. (iii) While public datasets for ASAS exist and are broadly used by researchers, the field is moving toward context-specific and language-specific ASAS datasets. There is also a need to train the ASAS model on nuanced adversarial and contrastive datasets. However, adversarial and contrastive datasets are considered scarce in the educational domain. Thus, there is potential for a study using RAG to augment the datasets **(RQ3)**. (iv) Current frameworks tend to use a hybrid approach. These methods combine neural models with handcrafted features or other architectures to improve accuracy. ASAS frameworks also face challenges in integration with the educational environment. Thus, there is a need for further study on how the ASAS framework can leverage recent technological advancements and be easily integrated into the educational environment **(RQ4)**. It is important to recognize the limitations of this review. The search was limited to papers published in English, which may have excluded important research published in other languages. Additionally, our analysis only included text-based assessments. This means the conclusions might not apply to scoring systems for math, graphics, or code-based answers. Because of the time it takes to publish studies, the most recent research may not be included in our analysis. Finally, significant technical challenges and ethical issues still need to be addressed, but the direction of ASAS is clear. The future of fair and effective educational

assessment will likely involve a blend of human skills and artificial intelligence, rather than the complete replacement of human educators. These systems should aim to enhance educators' abilities, enabling them to shift away from repetitive grading and toward the more complex, human-centered aspects of teaching and learning.

Acknowledgement

The authors would like to thank Brawijaya University for funding this research under the Adjunct Professor Grant with the registration number **04407/UN10.A0101/B/KS/2025**. The author also thanks the editor and anonymous reviewers for their contributions in improving the quality of this paper. The authors declared that Grammarly AI prompts were used to assist in grammar checking and language editing. All content generated was reviewed and verified by the authors, who take full responsibility for the final submission.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *W.H.N.P.;* ***data collection:*** *W.H.N.P.;* ***analysis and interpretation of results:*** *W.H.N.P., M.F.M.F.;* ***draft manuscript preparation:*** *W.H.N.P., M.F.M.F.;* ***writing-review and editing:*** *B.T.H., W.P., F.P., A.D.H.; All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Abdeljaber HA (2021) Automatic Arabic Short Answers Scoring Using Longest Common Subsequence and Arabic WordNet. IEEE Access 9:76433–76445. doi: 10.1109/ACCESS.2021.3082408

[2] Aghzal M, Bouni MAE, Driouech S, Mourhir A (2023) Compact Transformer-based Language Models for the Moroccan Darija. In: 2023 7th IEEE Congress on Information Science and Technology (CiSt). pp 299–304

[3] Agrawal S, Agrawal AJ (2019) Reviewing techniques for automatic response grading via language processing. Int J Innov Technol Explor Eng 8:1415–1420

[4] Akhilesh P, K AK, Bharadwaj SK, Venugopalan M (2024) Automated Short Answer Grading With Word Embedding-Based Semantic Similarity Using PySpark. In: 2024 IEEE Recent Advances in Intelligent Computational Systems (RAICS). pp 1–6

[5] Akila Devi TR, 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 Comput Sci 4. doi: 10.1007/s42979-023-01682-8

[6] Alhamed DH, Alajmi AM, Alqahtani TA, Alali YH, Alnassar MR, Alabbad DA (2022) iGrade: an automated short answer grading system. In: ACM International Conference Proceeding Series. pp 110–116

[7] Amur ZH, Hooi YK, Soomro GM (2022) Automatic Short Answer Grading (ASAG) using Attention-Based Deep Learning MODEL. In: 2022 International Conference on Digital Transformation and Intelligence (ICDI). pp 1–7

[8] Andersen N, Mang J, Goldhammer F, Zehner F (2025) Algorithmic Fairness in Automatic Short Answer Scoring. Int J Artif Intell Educ. doi: 10.1007/s40593-025-00495-5

[9] Ariely M, Salman A, Yarden A, Alexandron G (2025) Reflective prompt engineering: a new strategy for automated short answer scoring in biology. Int J Sci Educ 1–23. doi: 10.1080/09500693.2025.2523571

[10] Arifin AR, Purnamasari PD, Ratna AAP (2021) Automatic Essay Scoring for Indonesian Short Answers using Siamese Manhattan Long Short-Term Memory. In: 2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE). pp 1–6

[11] Asazuma Y, Funayama H, Matsubayashi Y, Mizumoto T, Reisert P, Inui K (2024) Take No Shortcuts! Stick to the Rubric: A Method for Building Trustworthy Short Answer Scoring Models. In: Communications in Computer and Information Science. pp 337–358

[12] Bai X, Stede M (2023) A Survey of Current Machine Learning Approaches to Student Free-Text Evaluation for Intelligent Tutoring. Int J Artif Intell Educ 33:992–1030. doi: 10.1007/s40593-022-00323-0

[13] Bani Saad M, Jackowska-Strumillo L, Bieniecki W (2025) Hybrid ANN-Based and Text Similarity Method for Automatic Short-Answer Grading in Polish. Appl Sci 15:1605. doi: 10.3390/app15031605

[14] Barlybayev A, Matkarimov B (2024) Development of system for generating questions, answers, distractors using transformers. Int J Electr Comput Eng 14:1851–1863. doi: 10.11591/ijece.v14i2.pp1851-1863

[15] Basak R, Naskar SK, Gelbukh A (2019) Short-answer grading using textual entailment. J Intell Fuzzy Syst 36:4909–4919. doi: 10.3233/JIFS-179038

[16] Basu S, Jacobs C, Vanderwende L (2013) Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading. Trans Assoc Comput Linguist 1:391–402. doi: 10.1162/tacl_a_00236

[17] Bexte M, Horbach A, Zesch T (2024) Strengths and weaknesses of automated scoring of free-text student answers. Inform-Spektrum. doi: 10.1007/s00287-024-01573-z

[18] Bhatt R, Patel M, Srivastava G, Mago V (2020) A Graph Based Approach to Automate Essay Evaluation. In: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, Toronto, ON, Canada, pp 4379–4385

[19] Blom B, Pereira JLM (2023) Domain Adaptation in Transformer Models: Question Answering of Dutch Government Policies. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp 196–208

[20] Bonthu S, Rama Sree S, Krishna Prasad MHM (2021) Automated Short Answer Grading Using Deep Learning: A Survey. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp 61–78

[21] Bonthu S, Rama Sree S, Krishna Prasad MHM (2023) Improving the performance of automatic short answer grading using transfer learning and augmentation. Eng Appl Artif Intell 123. doi: 10.1016/j.engappai.2023.106292

[22] Bonthu S, Sree SR, Krishna Prasad MHM (2024) Framework for automation of short answer grading based on domain-specific pre-training. Eng Appl Artif Intell 137. doi: 10.1016/j.engappai.2024.109163

[23] Bonthu S, Sree SR, Prasad MHMK (2024) SPRAG: building and benchmarking a Short Programming-Related Answer Grading dataset. Int J Data Sci Anal. doi: 10.1007/s41060-024-00576-z

[24] Camus L, Filighera A (2020) Investigating Transformers for Automatic Short Answer Grading. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp 43–48

[25] Carmon CM, Hu X, Graesser AC (2023) Assessment in Conversational Intelligent Tutoring Systems: Are Contextual Embeddings Really Better? In: Communications in Computer and Information Science. pp 121–129

[26] Carmon CM, Morgan B, Hu X, Graesser AC (2023) Automated Assessment of Initial Answers to Questions in Conversational Intelligent Tutoring Systems: Are Contextual Embedding Models Really Better? Electron Switz 12. doi: 10.3390/electronics12173654

[27] Casalino G, Cafarelli B, del Gobbo E, Fontanella L, Grilli L, Guarino A, Limone P, Schicchi D, Taibi D (2021) Framing automatic grading techniques for open-ended questionnaires responses. A short survey. In: CEUR Workshop Proceedings

[28] Chakraborty UK, Konar D, Roy S, Choudhury S (2019) Automatic short answer grading using rough concept clusters. Int J Adv Intell Paradig 14:260–280. doi: 10.1504/IJAIP.2019.103413

[29] Chamidah N, Santoni MM, Irmanda HN, Astriratma R, Tua LM, Yuniati T (2021) Word Expansion using Synonyms in Indonesian Short Essay Auto Scoring. In: 2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS. pp 296–300

[30] Chandrapati LM, Rao CK (2024) Descriptive Answers Evaluation Using Natural Language Processing Approaches. IEEE Access 12:87333–87347. doi: 10.1109/ACCESS.2024.3417706

[31] Chang L-H, Ginter F (2024) Automatic Short Answer Grading for Finnish with ChatGPT. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp 23173–23181

[32] Chang L-H, Kanerva J, Ginter F (2022) Towards Automatic Short Answer Assessment for Finnish as a Paraphrase Retrieval Task. In: BEA 2022 - 17th Workshop on Innovative Use of NLP for Building Educational Applications, Proceedings. pp 262–271

[33] Chang L-H, Rastas I, Pyysalo S, Ginter F (2021) Deep learning for sentence clustering in essay grading support. In: Proceedings of the 14th International Conference on Educational Data Mining, EDM 2021. pp 614–618

[34] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ n71. doi: 10.1136/bmj.n71

[35] Chaturvedi B, Basak R (2021) Automatic Short Answer Grading Using Corpus-Based Semantic Similarity Measurements. In: Advances in Intelligent Systems and Computing. pp 266–281

[36] Chavan SM, Prerana MS, Bathula R, Saikumar S, Dayalan G (2023) Automated Script Evaluation using Machine Learning and Natural Language Processing. In: 2023 2nd International Conference for Innovation in Technology, INOCON 2023

[37] 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. doi: 10.1109/ACCESS.2023.3260840

[38] Çınar A, Ince E, Gezer M, Yılmaz Ö (2020) Machine learning algorithm for grading open-ended physics questions in Turkish. Educ Inf Technol 25:3821–3844. doi: 10.1007/s10639-020-10128-0

[39] Clauser BE, Yaneva V, Baldwin P, An Ha L, Mee J (2024) Automated Scoring of Short-Answer Questions: A Progress Report. Appl Meas Educ 37:209–224. doi: 10.1080/08957347.2024.2386945

[40] Cohn C, Hutchins N, Le T, Biswas G (2024) A Chain-of-Thought Prompting Approach with LLMs for Evaluating Students' Formative Assessment Responses in Science. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp 23182–23190

[41] Condor A, Pardos Z, Linn M (2022) Representing Scoring Rubrics as Graphs for Automatic Short Answer Grading. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp 354–365

[42] De Gasperis G, Menini S, Tonelli S, Vittorini P (2019) Automated Grading of Short Text Answers: Preliminary Results in a Course of Health Informatics. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp 190–200

[43] del Gobbo E, Guarino A, Cafarelli B, Grilli L (2023) GradeAid: a framework for automatic short answers grading in educational contexts—design, implementation and evaluation. Knowl Inf Syst 65:4295–4334. doi: 10.1007/s10115-023-01892-9

[44] Divya A, Haridas V, Narayanan J (2023) Automation of Short Answer Grading Techniques: Comparative Study using Deep Learning Techniques. In: 2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT). pp 1–7

[45] Duong TNB, Meng CY (2024) Automatic Grading of Short Answers Using Large Language Models in Software Engineering Courses. In: 2024 IEEE Global Engineering Education Conference (EDUCON). pp 1–10

[46] Dzikovska M, Moore J, Steinhauser N, Campbell G, Farrow E, Callaway C (2010) BEETLE II: A system for tutoring and Computational Linguistics experimentation

[47] Dzikovska M, Nielsen R (2012) Towards effective tutorial feedback for explanation questions: a dataset and baselines

[48] Dzikovska MO, Nielsen RD, Leacock C (2016) The joint student response analysis and recognizing textual entailment challenge: making sense of student responses in educational applications. Lang Resour Eval 50:67–93. doi: 10.1007/s10579-015-9313-8

[49] Horbach A, Zesch T (2019) The Influence of Variance in Learner Answers on Automatic Content Scoring. Front Educ 4. doi: 10.3389/feduc.2019.

[50] Elnaka A, Nael O, Afifi H, Sharaf N (2021) AraScore: Investigating Response-Based Arabic Short Answer Scoring. In: Procedia CIRP. pp 282–291

[51] Fan C, Guo S, Wumaier A, Liu J (2023) A cross-attention and Siamese network based model for off-topic detection. In: 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI). pp 770–777

[52] Fateen M, Wang B, Mine T (2024) Beyond Scores: A Modular RAG-Based System for Automatic Short Answer Scoring With Feedback. IEEE Access 12:185371–185385. doi: 10.1109/ACCESS.2024.3508747

[53] Ferreira Mello R, Pereira Junior C, Rodrigues L, Pereira FD, Cabral L, Costa N, Ramalho G, Gasevic D (2025) Automatic Short Answer Grading in the LLM Era: Does GPT-4 with Prompt Engineering beat

Traditional Models? In: Proceedings of the 15th International Learning Analytics and Knowledge Conference. ACM, Dublin Ireland, pp 93–103

[54] Filighera A, Parihar SS, Ochs S, Steuer T, Meuser T (2022) Your Answer is Incorrect... Would you like to know why? Introducing a Bilingual Short Answer Feedback Dataset. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp 8577–8591

[55] Firoozi T, Gierl MJ (2024) Scoring Essays Written in Persian Using a Transformer-Based Model. In: The Routledge International Handbook of Automated Essay Evaluation, 1st ed. Routledge, New York, pp 55–77

[56] Funayama H, Asazuma Y, Matsubayashi Y, Mizumoto T, Inui K (2023) Reducing the Cost: Cross-Prompt Pre-finetuning for Short Answer Scoring. pp 78–89

[57] Funayama H, Matsubayashi Y, Asazuma Y, Mizumoto T, Inui K (2025) Cross-prompt Pre-finetuning of Language Models for Short Answer Scoring. Int J Artif Intell Educ. doi: 10.1007/s40593-025-00474-w

[58] Funayama H, Sato T, Matsubayashi Y, Mizumoto T, Suzuki J, Inui K (2022) Balancing Cost and Quality: An Exploration of Human-in-the-Loop Frameworks for Automated Short Answer Scoring. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp 465–476

[59] Gaikwad HR, Kiwelekar AW (2024) A Generative AI-Based Assistant to Evaluate Short and Long Answer Questions. SN Comput Sci 5. doi: 10.1007/s42979-024-02965-4

[60] Kotei E, Thirunavukarasu R (2023) A Systematic Review of Transformer-Based Pre-Trained Language Models through Self-Supervised Learning. Information 14:187. doi: 10.3390/info14030187

[61] Kavitha M, Akila K (2023) An Exploratory Study of Abstractive Text Summarization Using a Sequence-to-Sequence Model. In: 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS). pp 1–5

[62] Putnikovic M, Jovanovic J (2023) Embeddings for Automatic Short Answer Grading: A Scoping Review. IEEE Trans Learn Technol 16:219–231. doi: 10.1109/TLT.2023.3253071

[63] Xu W, Mahmud R, Lam Hoo W (2024) A Systematic Literature Review: Are Automated Essay Scoring Systems Competent in Real-Life Education Scenarios? IEEE Access 12:77639–77657. doi: 10.1109/ACCESS.2024.3399163

[64] Yang K, Raković M, Li Y, Guan Q, Gašević D, Chen G (2024) Unveiling the Tapestry of Automated Essay Scoring: A Comprehensive Investigation of Accuracy, Fairness, and Generalizability. Proc AAAI Conf Artif Intell 38:22466–22474. doi: 10.1609/aaai.v38i20.30254

[65] Grévisse C (2024) LLM-based automatic short answer grading in undergraduate medical education. BMC Med Educ 24. doi: 10.1186/s12909-024-06026-5

[66] Hahn MG, Navarro SMB, Valentín LDLF, Burgos D (2021) A Systematic Review of the Effects of Automatic Scoring and Automatic Feedback in Educational Settings. IEEE Access 9:108190–108198. doi: 10.1109/ACCESS.2021.3100890

[67] Hasanah U, Hartato BP (2020) Assessing Short Answers in Indonesian Using Semantic Text Similarity Method and Dynamic Corpus. In: 2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE). pp 312–316

[68] Hasanah U, Permanasari AE, Kusumawardani SS, Pribadi FS (2019) A scoring rubric for automatic short answer grading system. Telkomnika Telecommun Comput Electron Control 17:763–770. doi: 10.12928/TELKOMNIKA.V17I2.11785

[69] Henkel O, Hills L, Roberts B, McGrane J (2024) Can LLMs Grade Open Response Reading Comprehension Questions? An Empirical Study Using the ROARs Dataset. Int J Artif Intell Educ. doi: 10.1007/s40593-024-00431-z

[70] Horbach A, Pehlke J, Laarmann-Quante R, Ding Y (2024) Crosslingual Content Scoring in Five Languages Using Machine-Translation and Multilingual Transformer Models. Int J Artif Intell Educ 34:1294–1320. doi: 10.1007/s40593-023-00370-1

[71] 00028Ejima C, Takeuchi K (2022) Statistical Learning Models for Japanese Essay Scoring Toward One-shot Learning. In: 2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI). pp 313–318

[72] Ikiss S, Daoudi N, Abourezq M, Bellafkih M (2024) Improving Automatic Short Answer Scoring Task Through a Hybrid Deep Learning Framework. Int J Adv Comput Sci Appl 15:1066–1073. doi: 10.14569/IJACSA.2024.01508105

[73] Iqbal M, Udhiah RL, Nugraha TR, Pao H-K (2023) ASAGeR: Automated Short Answer Grading Regressor via Sentence Simplification. In: 2023 IEEE International Conference on Knowledge Graph (ICKG). pp 60–68

[74] Ivanova RV, Handschuh S (2024) Evaluating LLMs' Performance At Automatic Short-Answer Grading. In: CEUR Workshop Proceedings

[75] Jiang L, Bosch N (2024) Short answer scoring with GPT-4. In: L@S 2024 - Proceedings of the 11th ACM Conference on Learning @ Scale. pp 438–442

[76] Joseph N, Varghese SM (2025) Investigating the BERT Capabilities with GRU Model in Semantic Extraction for Short Answer Grading Tasks: A Regression Problem. In: 2025 2nd International Conference on Trends in Engineering Systems and Technologies (ICTEST). IEEE, Ernakulam, India, pp 1–5

[77] Kardam N, Misra S, Wilson D (2023) Is Natural Language Processing Effective in Education Research? A case study in student perceptions of TA support. In: ASEE Annual Conference and Exposition, Conference Proceedings

[78] Garg J, Papreja J, Apurva K, Jain G (2022) Domain-Specific Hybrid BERT based System for Automatic Short Answer Grading. In: 2022 2nd International Conference on Intelligent Technologies (CONIT). pp 1–6

[79] Kaya M, Cicekli I (2024) A Hybrid Approach for Automated Short Answer Grading. IEEE Access 12:96332–96341. doi: 10.1109/ACCESS.2024.3420890

[80] Kazi N, Kahanda I (2023) Enhancing Transfer Learning of LLMs through Fine- Tuning on Task - Related Corpora for Automated Short-Answer Grading. In: 2023 International Conference on Machine Learning and Applications (ICMLA). pp 1687–1691

[81] Ke Z, Ng V (2019) Automated Essay Scoring: A Survey of the State of the Art. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, Macao, China, pp 6300–6308

[82] Li C, Lin L, Mao W, Xiong L, Lin Y (2022) An Automated Essay Scoring model Based on Stacking Method. In: 2022 IEEE 2nd International Conference on Software Engineering and Artificial Intelligence (SEAI). IEEE, Xiamen, China, pp 248–252

[83] Gao R, Thomas N, Srinivasa A (2023) Work in Progress: Large Language Model Based Automatic Grading Study. In: 2023 IEEE Frontiers in Education Conference (FIE). pp 1–4

[84] Kripalani S, Lipare T, Kadam D (2024) Subjective Answer Grader using Semantic Similarity and Keyword Matching. In: 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). pp 1–7

[85] Kumar AP, Nayak A, Shenoy MK, Manoj RJ, Priyadarshi A (2022) Pattern-Based Syntactic Simplification of Compound and Complex Sentences. IEEE Access 10:53290–53306. doi: 10.1109/ACCESS.2022.3174846

[86] Kwok-Fai Lui A, Ng S-C, Wing-Nga Cheung S (2024) Automated short answer grading with computer-assisted grading example acquisition based on active learning. Interact Learn Environ 32:2087–2104. doi: 10.1080/10494820.2022.2137530

[87] Lakshmi PS, Simha JB, Ranjan R (2024) IntelliGrader: A Framework for Automatic Short Answer Grading, Inconsistency Check and Feedback in Educational Context-conception, Implementation and Evaluation. Karbala Int J Mod Sci 10:449–461. doi: 10.33640/2405-609X.3370

[88] Lalita Luhurkinanti D, Dewi Purnamasari P, Tsunakawa T, Agung Putri Ratna A (2025) Japanese Short Answer Grading for Japanese Language Learners Using the Contextual Representation of BERT. IEEE Access 13:17195–17207. doi: 10.1109/ACCESS.2025.3532659

[89] Kortemeyer G (2024) Performance of the pre-trained large language model GPT-4 on automated short answer grading. Discov Artif Intell 4. doi: 10.1007/s44163-024-00147-y

[90] Zhou X, Fan X, Yang Y, Ren G (2022) Attention Word Embedding Network-Based Lightweight Automatic Essay Scoring Model. In: Liang Q, Wang W, Mu J, Liu X, Na Z (eds) Artificial Intelligence in China. Springer Singapore, Singapore, pp 275–282

[91] Zhou X, Yang L, Fan X, Ren G, Yang Y, Lin H (2021) Self-training vs Pre-trained Embeddings for Automatic Essay Scoring. In: Lin H, Zhang M, Pang L (eds) Information Retrieval. Springer International Publishing, Cham, pp 155–167

[92] Zhu X, Wu H, Zhang L (2022) Automatic Short-Answer Grading via BERT-Based Deep Neural Networks. IEEE Trans Learn Technol 15:364–375. doi: 10.1109/TLT.2022.3175537

[93] Maslim M, Wang H-C, Putra CD, Prabowo YD (2024) A Trustworthy Automated Short-Answer Scoring System Using a New Dataset and Hybrid Transfer Learning Method. Int J Interact Multimed Artif Intell 8:37–45. doi: 10.9781/ijimai.2024.02.003

[94] Meccawy M, Bayazed AA, Al-Abdullah B, Algamdi H (2023) Automatic Essay Scoring for Arabic Short Answer Questions using Text Mining Techniques. Int J Adv Comput Sci Appl 14:768–775. doi: 10.14569/IJACSA.2023.0140682

[95] Mizumoto T, Ouchi H, Isobe Y, Reisert P, Nagata R, Sekine S, Inui K (2019) Analytic Score Prediction and Justification Identification in Automated Short Answer Scoring

[96] Mohler M, Bunescu R, Mihalcea R (2011) Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In: Lin D, Matsumoto Y, Mihalcea R (eds) Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Portland, Oregon, USA, pp 752–762

[97] Mountantonakis M, Mertzanis L, Bastakis M, Tzitzikas Y (2024) A comparative evaluation for question answering over Greek texts by using machine translation and BERT. Lang Resour Eval. doi: 10.1007/s10579-024-09745-9

[98] Muhammad A-R, Permanasari AE, Hidayah I (2022) Enhancing GAN-LCS Performance Using an Abbreviations Checker in Automatic Short Answer Scoring. Computers 11. doi: 10.3390/computers11070108

[99] Mukti AAS, Alfarozi SAI, Kusumawardani SS (2023) Transformers Based Automated Short Answer Grading with Contrastive Learning for Indonesian Language. In: 2023 15th International Conference on Information Technology and Electrical Engineering (ICITEE). pp 133–138

[100] Nampoothiri AbhijithN, Thushara MG, N Pradeesh, R AKK, Santhosh A, S AK, Kumar S (2025) The Impact of Transformer Models and Regression Algorithms on Automated Short Answer Grading. In: 2025 Fifth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT). IEEE, Bhilai, India, pp 1–6

[101] Ndukwe IG, Amadi CE, Nkomo LM, Daniel BK (2020) Automatic Grading System Using Sentence-BERT Network. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp 224–227

[102] Nielsen R, Ward W, Martin J, Palmer M (2008) Annotating Students' Understanding of Science Concepts

[103] Oka H, Nguyen HT, Nguyen CT, Nakagawa M, Ishioka T (2022) Fully Automated Short Answer Scoring of the Trial Tests for Common Entrance Examinations for Japanese University. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp 180–192

[104] Oktaviani AN, Alief MZ, Santiar L, Purnamasari PD, Ratna AAP (2021) Automatic Essay Grading System for Japanese Language Exam using CNN-LSTM. In: 2021 17th International Conference on Quality in Research (QIR): International Symposium on Electrical and Computer Engineering. pp 164–169

[105] Olowolayemo A, Nawi SD, Mantoro T (2019) Short answer scoring in english grammar using text similarity measurement. In: Proceedings - 2018 4th International Conference on Computing, Engineering, and Design, ICCED 2018. pp 131–136

[106] Ormerod C, Lottridge S, Harris AE, Patel M, van Wamelen P, Kodeswaran B, Woolf S, Young M (2023) Automated Short Answer Scoring Using an Ensemble of Neural Networks and Latent Semantic Analysis Classifiers. Int J Artif Intell Educ 33:467–496. doi: 10.1007/s40593-022-00294-2

[107] Osaka J, Maeda A, Oka H, Mori Y, Ishioka T, Suyari H (2025) Reliable and efficient automated short-answer scoring for a large dataset using active learning and deep learning. Interact Learn Environ 1–12. doi: 10.1080/10494820.2025.2452005

[108] Ouahrani L, Bennouar D (2020) AR-ASAG An ARabic Dataset for Automatic Short Answer Grading Evaluation

[109] Ouahrani L, Bennouar D (2024) Paraphrase Generation and Supervised Learning for Improved Automatic Short Answer Grading. Int J Artif Intell Educ 34:1627–1670. doi: 10.1007/s40593-023-00391-w

[110] P SL, B SJ (2022) A Hybrid Qualitative and Quantitative Approach for Automatic Short Answer Grading Using Classification Algorithms. In: 2022 4th International Conference on Circuits, Control, Communication and Computing (I4C). pp 12–17

[111] Padó U, Eryilmaz Y, Kirschner L (2024) Short-Answer Grading for German: Addressing the Challenges. Int J Artif Intell Educ 34:1321–1352. doi: 10.1007/s40593-023-00383-w

[112] Charitopoulos A, Rangoussi M, Metafas D, Koulouriotis D (2025) Text mining technologies applied to free-text answers of students in e-assessment. Discov Comput 28:5. doi: 10.1007/s10791-024-09496-9

[113] Petricioli L, Skračić K, Petrović J, Pale P (2023) Exploring Pre-scoring Clustering for Short Answer Grading. In: 2023 46th MIPRO ICT and Electronics Convention (MIPRO). pp 1567–1571

[114] Prabhudesai A, Duong TNB (2019) Automatic Short Answer Grading using Siamese Bidirectional LSTM Based Regression. In: 2019 IEEE International Conference on Engineering, Technology and Education (TALE). pp 1–6

[115] Ghavidel HA, Zouaq A, Desmarais MC (2020) Using BERT and XLNET for the automatic short answer grading task. In: CSEDU 2020 - Proceedings of the 12th International Conference on Computer Supported Education. pp 58–67

[116] Putri Ratna AA, Kaltsum A, Santiar L, Khairunissa H, Ibrahim I, Purnamasari PD (2019) Term Frequency-Inverse Document Frequency Answer Categorization with Support Vector Machine on Automatic Short Essay Grading System with Latent Semantic Analysis for Japanese Language. In: ICECOS 2019 - 3rd International Conference on Electrical Engineering and Computer Science, Proceeding. pp 293–298

[117] Qiao C, Hu X (2023) Leveraging Semantic Facets for Automatic Assessment of Short Free Text Answers. IEEE Trans Learn Technol 16:26–39. doi: 10.1109/TLT.2022.3199469

[118] Qiu H, White B, Ding A, Costa R, Hachem A, Ding W, Chen P (2024) SteLLA: A Structured Grading System Using LLMs with RAG. In: 2024 IEEE International Conference on Big Data (BigData). pp 8154–8163

[119] Rababah H, Al-Taani AT (2017) An automated scoring approach for Arabic short answers essay questions. In: 2017 8th International Conference on Information Technology (ICIT). IEEE, Amman, Jordan, pp 697–702

[120] Rahutomo F, Ari Roshinta T, Rohadi E, Siradjuddin I, Ariyanto R, Setiawan A, Adhisuwignjo S (2018) Open Problems in Indonesian Automatic Essay Scoring System. Int J Eng Technol 7:156. doi: 10.14419/ijet.v7i4.44.26974

[121] Raja Subramanian R, Surekha B, SubbaRao P, Kesava Chowdary GA, Venkatesh M (2025) EduGuard – AI Based Online Exam Management System. In: 2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE). IEEE, Virudhunagar, India, pp 1–6

[122] Rajpurkar P, Jia R, Liang P (2018) Know What You Don't Know: Unanswerable Questions for SQuAD

[123] Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuAD: 100,000+ Questions for Machine Comprehension of Text

[124] Ramesh D, Sanampudi SK (2023) Semantic and Linguistic Based Short Answer Scoring System. Int J Intell Syst Appl Eng 11:246–251

[125] Ramesh D, Sanampudi SK (2024) Coherence-based automatic short answer scoring using sentence embedding. Eur J Educ 59. doi: 10.1111/ejed.12684

[126] Ramnarain-Seetohul V, Bassoo V, Rosunally Y (2022) Work-in-Progress: Computing Sentence Similarity for Short Texts using Transformer models. In: 2022 IEEE Global Engineering Education Conference (EDUCON). pp 1765–1768

[127] Ramnarain-Seetohul V, Bassoo V, Rosunally Y (2023) Obmaaq: Ontology-Based Model for Automated Assessment of Short-Answer Questions. In: 2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI). pp 1–8

[128] Ran Y, He B, Xu J (2018) A Study on Performance Sensitivity to Data Sparsity for Automated Essay Scoring. In: Liu W, Giunchiglia F, Yang B (eds) Knowledge Science, Engineering and Management. Springer International Publishing, Cham, pp 104–116

[129] Ratna AAP, Khairunissa H, Kaltsum A, Ibrahim I, Purnamasari PD (2019) Automatic Essay Grading for Bahasa Indonesia with Support Vector Machine and Latent Semantic Analysis. In: 2019 International Conference on Electrical Engineering and Computer Science (ICECOS). pp 363–367

[130] Ratna AAP, Noviaindriani RR, Santiar L, Ibrahim I, Purnamasari PD (2019) K-Means Clustering for Answer Categorization on Latent Semantic Analysis Automatic Japanese Short Essay Grading System. In: 2019 16th International Conference on Quality in Research (QIR): International Symposium on Electrical and Computer Engineering. pp 1–5

[131] Ratna AAP, Wulandari NA, Kaltsum A, Ibrahim I, Purnamasari PD (2019) Answer Categorization Method Using K-Means for Indonesian Language Automatic Short Answer Grading System Based on Latent Semantic Analysis. In: 2019 16th International Conference on Quality in Research (QIR): International Symposium on Electrical and Computer Engineering. pp 1–5

[132] Reshmi SN, Shreelekshmi R (2019) Textual Entailment based on Semantic Similarity Using WordNet. In: 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT). pp 1188–1192

[133] Runyon CR, Paniagua MA, Rosenthal FA, Veneziano AL, McNaughton L, Murray CT, Harik P (2024) SHARP (SHort Answer, Rationale Provision): A New Item Format to Assess Clinical Reasoning. Acad Med 99:976–980. doi: 10.1097/ACM.0000000000005769

[134] Sahu A, Bhowmick PK (2020) Feature Engineering and Ensemble-Based Approach for Improving Automatic Short-Answer Grading Performance. IEEE Trans Learn Technol 13:77–90. doi: 10.1109/TLT.2019.2897997

[135] Salam MA, El-Fatah MA, Hassan NF (2022) Automatic grading for Arabic short answer questions using optimized deep learning model. PLoS ONE 17. doi: 10.1371/journal.pone.0272269

[136] Salim HR, De C, Pratamaputra ND, Suhartono D (2022) Indonesian automatic short answer grading system. Bull Electr Eng Inform 11:1586–1603. doi: 10.11591/eei.v11i3.3531

[137] Sanuvala G, Fatima SS, Kambhampati T, Sanuvala R (2024) Automatic Short Answer Scoring on an Indian Dataset Using Transformer-Based Language Models. In: Lecture Notes in Networks and Systems. pp 287–295

[138] Sanz-Fayos J, de-la-Fuente-Valentín L, Verdú E (2022) Keyword-Based Processing for Assessing Short Answers in the Educational Field. In: Communications in Computer and Information Science. pp 134–146

[139] Saskia L, Hidayah I, Kusumawardani SS (2022) Improvement of GAN-LCS Performance with Synonym Recognition. In: 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE). pp 40–45

[140] Schleifer AG, Klebanov BB, Ariely M, Alexandron G (2023) Transformer-based Hebrew NLP models for Short Answer Scoring in Biology. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Toronto, ON, Canada, pp 550–555

[141] Schneider J, Schenk B, Niklaus C (2024) Towards LLM-Based Autograding for Short Textual Answers. In: International Conference on Computer Supported Education, CSEDU - Proceedings. pp 280–288

[142] Seneviratne HMTW, Manathunga SS (2025) Artificial intelligence assisted automated short answer question scoring tool shows high correlation with human examiner markings. BMC Med Educ 25:1146. doi: 10.1186/s12909-025-07718-2

[143] Shah N, Thakkar HK, Mewada H (2024) On the Analysis of a BERT-based Domain-Specific Question Answering Models for Indian Legal System. In: 2024 6th International Symposium on Advanced Electrical and Communication Technologies (ISAECT). pp 1–7

[144] Sharmila P, Anbananthen KSM, Chelliah D, Parthasarathy S, Balasubramaniam B, Lurudusamy SN (2024) Transformer-Based Sequence Modeling Short Answer Assessment Framework. HighTech Innov J 5:627–639. doi: 10.28991/hij-2024-05-03-06

[145] Shermis MD (2015) Contrasting State-of-the-Art in the Machine Scoring of Short-Form Constructed Responses. Educ Assess 20:46–65. doi: 10.1080/10627197.2015.997617

[146] Shylesh A, Raafeh A, Mathin S, Prakash VB, Shanmugasundaram H (2023) Automated Answer Script Evaluation Using Deep Learning. In: 2023 International Conference on Computer Communication and Informatics (ICCCI). pp 1–5

[147] Sirotheau S, dos Santos JCA, Favero EL, de Freitas SN (2019) Automated evaluation of short answers using text similarity for the Portuguese language. J Comput Sci 15:1669–1677. doi: 10.3844/jcssp.2019.1669.1677

[148] Situmeang SIG, Sihite RMGT, Simanjuntak H, Amalia J (2023) A Deep Learning-Based Regression Approach to Indonesian Short Answer Grading System. In: ACM International Conference Proceeding Series. pp 201–209

[149] Somers R, Cunningham-Nelson S, Boles W (2021) Applying natural language processing to automatically assess student conceptual understanding from textual responses. Australas J Educ Technol 37:98–115. doi: 10.14742/ajet.7121

[150] Soulimani YA, El Achaak L, Bouhorma M (2024) Deep learning based Arabic short answer grading in serious games. Int J Electr Comput Eng 14:841–853. doi: 10.11591/ijece.v14i1.pp841-853

[151] Sree Lakshmi P, Simha JB, Ranjan R (2024) Empowering Educators: Automated Short Answer Grading with Inconsistency Check and Feedback Integration using Machine Learning. SN Comput Sci 5. doi: 10.1007/s42979-024-02954-7

[152] Sung C, Ma T, Dhamecha TI, Reddy V, Saha S, Arora R (2019) Pre-training BERT on domain resources for short answer grading. In: EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference. Association for Computational Linguistics, Hong Kong, pp 6071–6075

[153] Surya K, Gayakwad E, Nallakaruppan MK (2019) Deep learning for short answer scoring. Int J Recent Technol Eng 7:1712–1715

[154] Takano S, Ichikawa O (2022) Automatic scoring of short answers using justification cues estimated by BERT. In: BEA 2022 - 17th Workshop on Innovative Use of NLP for Building Educational Applications, Proceedings. Association for Computational Linguistics (ACL), Seattle, pp 8–13

[155] Tan H, Wang C, Duan Q, Lu Y, Zhang H, Li R (2023) Automatic short answer grading by encoding student responses via a graph convolutional network. Interact Learn Environ 31:1636–1650. doi: 10.1080/10494820.2020.1855207

[156] Tang X, Chen H, Lin D, Li K (2024) Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. Heliyon 10:e34262. doi: 10.1016/j.heliyon.2024.e34262

[157] Tareaf RB (2022) MBTI BERT: A Transformer-Based Machine Learning Approach Using MBTI Model For Textual Inputs. In: 2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys). pp 2285–2292

[158] Tashu TM, Horváth T (2020) SmartScore-Short Answer Scoring Made Easy Using Sem-LSH. In: 2020 IEEE 14th International Conference on Semantic Computing (ICSC). pp 145–149

[159] Törnqvist M, Mahamud M, Guzman EM, Farazouli A (2023) ExASAG: Explainable Framework for Automatic Short Answer Grading. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp 361–371

[160] Tulu CN, Ozkaya O, Orhan U (2021) Automatic Short Answer Grading With SemSpace Sense Vectors and MaLSTM. IEEE Access 9:19270–19280. doi: 10.1109/ACCESS.2021.3054346

[161] Uto M (2023) Neural Automated Short-Answer Grading Considering Examinee-Specific Features. In: 2023 IEEE International Conference on Advanced Learning Technologies (ICALT). pp 336–338

[162] Uto M, Uchida Y (2020) Automated Short-Answer Grading Using Deep Neural Networks and Item Response Theory. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp 334–339

[163] de-la-Fuente-Valentín L, Verdú E, Padilla-Zea N, Villalonga C, Blanco Valencia XP, Baldiris Navarro SM (2022) Semiautomatic Grading of Short Texts for Open Answers in Higher Education. In: Communications in Computer and Information Science. pp 49–62

[164] Wahyuningsih T (2021) Text Mining an Automatic Short Answer Grading (ASAG), Comparison of Three Methods of Cosine Similarity, Jaccard Similarity and Dice's Coefficient. J Appl Data Sci 2:45–54. doi: 10.47738/jads.v2i2.31

[165] Wang B, Dawton B, Ishioka T, Mine T (2024) Optimizing Answer Representation Using Metric Learning for Efficient Short Answer Scoring. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp 236–248

[166] Wang J, Chen J, Ou X, Han Q, Tang Z (2023) Multi-level Feature Fusion for Automated Essay Scoring. J Netw Intell 8:76–88

[167] Wang Z, Lan AS, Waters AE, Grimaldi P, Baraniuk RG (2019) A meta-learning augmented bidirectional transformer model for automatic short answer grading. In: EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining. pp 667–670

[168] Wangwiwattana C, Tongvivat Y (2022) Semi-automatic short answers clustering and grading with K-Means and Keyword Matching algorithms. In: 2022 6th International Conference on Information Technology (InCIT). pp 280–284

[169] Wijanto MC, Yong H-S (2024) Combining Balancing Dataset and SentenceTransformers to Improve Short Answer Grading Performance. Appl Sci Switz 14. doi: 10.3390/app14114532

[170] Wijaya MC (2021) Automatic Short Answer Grading System in Indonesian Language Using BERT Machine Learning. Rev Intell Artif 35:503–509. doi: 10.18280/ria.350609

[171] Wilianto D, Girsang AS (2023) Automatic Short Answer Grading on High School's E-Learning Using Semantic Similarity Methods. TEM J 12:297–302. doi: 10.18421/TEM121-37

[172] Wiratmo A, Fatichah C (2020) Assessment of Indonesian Short Essay using Transfer Learning Siamese Dependency Tree-LSTM. In: 2020 4th International Conference on Informatics and Computational Sciences (ICICoS). pp 1–5

[173] Wu M, Li S, Teng K, Du C, Duan J (2021) Investigations on Answer Selection Based on Co-attention Guided Stacked BiLSTM with BERT. In: 2021 IEEE Sixth International Conference on Data Science in Cyberspace (DSC). pp 420–427

[174] Wu S-H, Yeh C-Y (2019) A Short Answer Grading System in Chinese by CNN. In: 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST). pp 1–5

[175] Goenka P, Piplani M, Sawhney R, Mathur P, Shah RR (2020) ESAS: Towards Practical and Explainable Short Answer Scoring. In: AAAI 2020 - 34th AAAI Conference on Artificial Intelligence. pp 13797–13798

[176] Yan Z, Zhang R, Jia F (2024) Exploring the Potential of Large Language Models as a Grading Tool for Conceptual Short-Answer Questions in Introductory Physics. In: ACM International Conference Proceeding Series. pp 308–314

[177] Gomaa WH, Nagib AE, Saeed MM, Algarni A, Nabil E (2023) Empowering Short Answer Grading: Integrating Transformer-Based Embeddings and BI-LSTM Network. Big Data Cogn Comput 7. doi: 10.3390/bdcc7030122

[178] Yang R, Cao J, Wen Z, Shen J (2022) Automated post scoring: evaluating posts with topics and quoted posts in online forum. World Wide Web 25:1197–1221. doi: 10.1007/s11280-022-01005-6

[179] Yeruva N, Venna S, Indukuri H, Marreddy M (2022) Triplet Loss based Siamese Networks for Automatic Short Answer Grading. In: ACM International Conference Proceeding Series. pp 60–64

[180] Zeng Z, Gašević D, Chen G (2023) On the Effectiveness of Curriculum Learning in Educational Text Scoring. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023. pp 14602–14610

[181] Zhang L, Huang Y, Yang X, Yu S, Zhuang F (2022) An automatic short-answer grading model for semi-open-ended questions. Interact Learn Environ 30:177–190. doi: 10.1080/10494820.2019.1648300

[182] Zhang M, Baral S, Heffernan N, Lan A (2022) Automatic Short Math Answer Grading via In-context Meta-learning. In: Proceedings of the 15th International Conference on Educational Data Mining, EDM 2022

[183] Li Z, Tomar Y, Passonneau RJ (2021) A Semantic Feature-Wise Transformation Relation Network for Automatic Short Answer Grading. In: EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings. pp 6030–6040

[184] Lubis FF, Putri A, Waskita D, Sulistyaningtyas T, Arman AA, Rosmansyah Y (2021) Automated Short-Answer Grading using Semantic Similarity based on Word Embedding. Int J Technol 12:571–581. doi: 10.14716/ijtech.v12i3.4651

[185] Luo L, Yang H, Li Z, Pedrycz W (2024) Learning to Score: A Coding System for Constructed Response Items via Interactive Clustering. Systems 12. doi: 10.3390/systems12090380