Survey and Analysis for the Challenges in Computer Science to the Automation of Grading Systems

JOAN LU, University of Huddersfield, Huddersfield, United Kingdom of Great Britain and Northern Ireland
BHAVYA KRISHNA BALASUBRAMANIAN, Computer Science, University of Huddersfield, Huddersfield, United Kingdom of Great Britain and Northern Ireland and Computer Science, Elizabeth School of London, London, United Kingdom
MIKE JOY, Computer Science, University of Warwick, Coventry, United Kingdom of Great Britain and Northern Ireland
QIANG XU, Computing and Engineering, University of Huddersfield, Huddersfield, United Kingdom of Great Britain and Northern Ireland

Assessment is essential to educational system. Automatic grading reduces the time and effort taken by tutors to assess the answers written by the students. To understand recent computational methods used for automatic grading, a review has been conducted. 4,084 articles were initially identified using a keyword search. After filtering, the number was reduced to 57. It was found that statistical models are normally used in Automatic-Short-Answer-Grading (ASAG); vector-based similarity measures are the most popular among projects; pilot datasets are mostly used; standard datasets for evaluation are missing. Evidence shows that machine learning and deep learning are most popularly adopted methods and generative AI, e.g., LLMs and ChatGPT are also jump to the chance, which indicates that integrating AI in education is an inevitable trend. Also, most investigations prefer to adopt multiple approaches to improve computational quality, dataset analysis, and evaluation results. The identified research gaps will be a useful reference guide to users/researchers beneficial to formative/summative assessment. We concluded that the presented outcome, analysis and discussions are informative to academia and pedagogical practitioners who are interested in further developing/using ASAG systems. Although research into ASAG is still rudimentary, it is a promising area with impact on academic circles/commercially educational markets.

CCS Concepts: • **Applied computing;** • **Computing methodologies;** • **Information systems**;

Additional Key Words and Phrases: Formative/Summative Assessment, Intelligence learning, grading systems, embedding, machine learning, deep learning, natural language processing, generative AI, ChatGPT, LLMs

This work was partially sponsored by the PhD's scholarship from the University of Huddersfield.
Authors' Contact Information: Joan Lu, University of Huddersfield, Huddersfield, Kirklees, United Kingdom of Great Britain and Northern Ireland; e-mail: Z.Lu@leedsbeckett.ac.uk; Bhavya Krishna Balasubramanian, Computer Science, University of Huddersfield, Huddersfield, West Yorkshire, United Kingdom of Great Britain and Northern Ireland and Computer Science, Elizabeth School of London, London, London, United Kingdom; e-mail: rajabhavya28@gmail.com; Mike Joy, Computer Science, University of Warwick, Coventry, West Midlands, United Kingdom of Great Britain and Northern Ireland; e-mail: m.s.joy@warwick.ac.uk; Qiang Xu, Computing and Engineering, University of Huddersfield, Huddersfield, Kirklees, United Kingdom of Great Britain and Northern Ireland; e-mail: q.xu2@hud.ac.uk.
Author's Current Contact Information: Joan Lu, Leeds Beckett University, Leeds, United Kingdom of Great Britain and Northern Ireland.

ACM 0360-0300/2025/09-ART3
https://doi.org/10.1145/3748521

ACM Reference Format:
Joan Lu, Bhavya Krishna Balasubramanian, Mike Joy, and Qiang Xu. 2025. Survey and Analysis for the Challenges in Computer Science to the Automation of Grading Systems. *ACM Comput. Surv.* 58, 1, Article 3 (September 2025), 37 pages. https://doi.org/10.1145/3748521

1 Introduction

A survey report of researchandmarkets.com indicates that the global intelligence in education market is forecast to grow from USD 537m in 2018 to USD 3,683m by 2022 [1]. With the increasing adoption of advanced techniques in education, the impacts of intelligent learning, such as smart content, personalisation, virtual lectures, auto assessment, and virtual learning environments, will enhance further in near future [2]. Assessment is one of essential components in any educational system that affects the curriculum, student performance, and teaching methods [3, 4].

Automatic short answer grading – ASAG can be used for both formative and summative assessment depending on the pedagogical purpose. Recently, it has been reported that ASAG has been used into the auto examination systems when the assessment is associated with descriptive answers [5–8]. If the questions are descriptive in nature, there may be multiple explanations or multiple answers that may be of value to guide the learners, e.g., questions used in social science [9–11]. Thus, guidance may be a purpose of formative assessment. If a question is of a quantitative nature, summative assessment is presented as student's records.

This article will cover ASAG with the relationship of formative and summative assessment [9–12], decision-making for text similarities [13–15], dataset availability [16–19], computational accuracy [13, 14], and evaluation methods for existing ASAG systems. Based on the collected and analysed data, we believe that well-developed ASAG tools will make a positive educational impact that when ASAG is used to support pedagogical practice.

2 Conducted Reviewing Methods

We look for the literature on the ASAG related approaches to understand what the current " state-of-the-art" is, from both technological and educational perspectives. The following **research questions** (**RQs**) are addressed in this study.

- **RQ1.** What existing assessment methods and technologies are associated with ASAG? It will cover a wide range of methods relevant to the educational practice and reported technologies that are adopted or still at the stage of research labs.
- **RQ2.** What computational technical methods and modes of similarity measures are currently used by ASAG? It needs to understand the challenges in computational modelling with different approaches and attempts, especially with natural language processing and AI related technologies, such as machine learning and deep learning methods.
- **RQ3.** How short answers are graded in existing ASAG tools? It needs to investigate the technology readiness and what are missed if the technology is adopted in the real situation.
- **RQ4.** What are the existing pedagogical challenges to evaluating short answers automatically? It needs to hear what users say about their experience of using the technology.
- **RQ5.** What are the possible ways to improve short answer scoring systems? It needs to make recommendations to help the decision making for both academic users and technology developers.

To answer the above questions, we reviewed articles chosen from the following academic search engines: IEEE Explore, sciencedirect.com, Scopus, and Springer. We used the search terms, such as

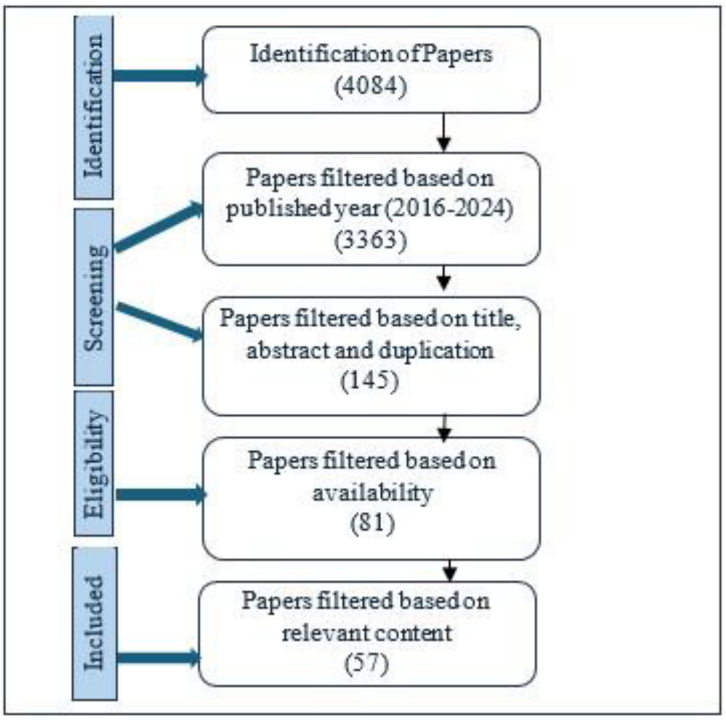

Fig. 1. Research article selection process.

"short answer grading" OR "short answer scoring" OR "open text grading" OR "free text grading" OR "automated grading". The process of selecting the research articles began by identifying with input keywords; followed by filtering the relevance to the topic of ASAG, then removing duplicate articles, and finally keeping the available articles ready to review.

We conducted the review using PRISMA method [20]. Initially we identified 4,084 articles from the above-mentioned publication databases. After filtering the articles from the years 2016 to 2024, we reduced the samples to 3,363 articles by analysing the titles and abstracts. Meanwhile, all the duplicate articles were eliminated, and we ended up with 145 articles. We checked the filtered articles for the availability of full text and removed ones without full text from the selection list, reducing the list of articles to 81. Finally, we ended up with 57 articles by completely reading the whole context and selecting only the relevant articles. The process of article selection is shown in Figure 1.

Evidence shows in Figure 2 and Table 1 covering the period 2016-2024. After removing duplicated publications from different search engines, we found that the most articles are indexed in sciencedirect.com / Scopus, about 2,493 articles. The second place is Springer, about 813 articles, and then about 57 articles in IEEE Explorer. Within the total number of 3,363 articles, the annual number of publications increased 4-fold from 142 in 2016 to 735 in 2024. The scope will cover the challenges of ASAG in grading methods [12], decision-making for text similarities [13–15], dataset availability [16–19], computational accuracy and evaluation methods [13, 14]. Based on the collected and analysed facts, we found that research into ASAG is growing and believe that well-developed ASAG tools will make a positive impact on education institutes.

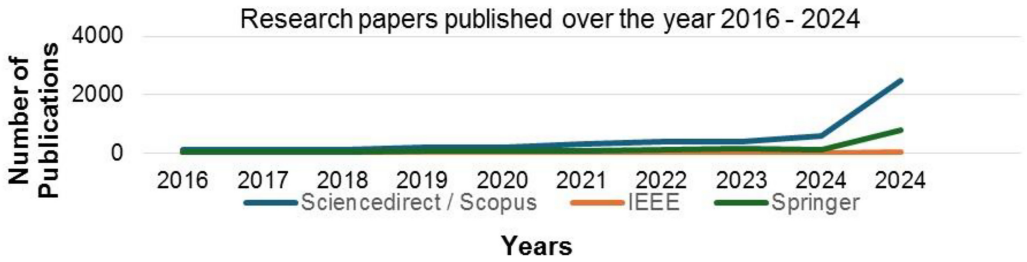

Fig. 2. ASAG Research articles published over the year 2016-2024(accessed November 2024).

Table 1. ASAG Research Articles Published in ScienceDirect / Scopus, IEEE Explorer, and Springer

Online Repository	Total	Year								
		2016	2017	2018	2019	2020	2021	2022	2023	2024
ScienceDirect / Scopus	2,493	108	131	136	188	204	328	402	410	586
IEEE	57	2	2	11	7	8	4	11	5	7
Springer	813	32	54	48	79	88	90	118	162	142
	3,363	142	187	195	274	300	422	531	577	735

3 Pedagogical Assessment Methods with ASAG

Pedagogically, there are two forms of assessments are commonly implemented in educational institutes, i.e., formative assessment and summative assessment. ASAG is associated with both assessment forms.

—**Formative assessment** has a positive impact on classroom learning. The learning pattern is process based rather than outcome based [21]. Lu et al reported several case studies of using a wireless response system – WRS in the classroom teaching or industrial training [9–11, 22]. During delivering the lecture, the teacher presented a quiz to test the students' understanding through the WRS. Then, students responded via WRS interactively either individually or in a group after discussing the questions with the teacher or their peers. The ASAG embedded in WRS could provide indicative grades for the cohort. The purpose is to measure whether the knowledge has been grasped or delivered thoroughly at the session. The teacher could immediately observe the effectiveness of teaching and learning [9, 10]. Therefore, the teaching pattern or speed can be adjusted into an acceptable pace that suits learners' cognition and curriculum requirements. Good examples have been shown in industrial training and primary school learning, and the approach has been shown to be effective at different educational levels [9–11]. Thus, formative assessment can be considered as being associated with summative assessment [9, 10, 22]. Broadbent et al. also reported another benefit of formative assessment, namely that it can improve **self-regulated learning** (**SRL**), which can help students to prepare what should be done to achieve the expected summative outcome, because the learners have benefited the feedback during the learning process [21]. With formative assessment, students can have opportunity to improve or adjust their learning plans and strategies after initial thoughts or ideas before the final assessment. With a good design of formative assessment, the outcome of summative assessment can be subsequently improved, and ultimately the retention and progression rates are improved significantly as well [9, 10].

—**Summative assessment** is another assessment form as a part of pedagogical curriculum design in most educational institutes, when the student performance is officially measured

Table 2. Formative assessment vs. Summative assessment

No	Formative Assessment	Summative Assessment
1.	Used to assess the student's continuous development throughout the course	Used to assess the students overall understanding at the end of the course
2.	Useful for the tutor to modify the lesson plan depending on the assessment evaluation	Useful to assess the tutors' practice and motivates student's accountability
3.	Examples: class work, homework and quizzes	Examples: end of term exam, SAT, benchmark assessment

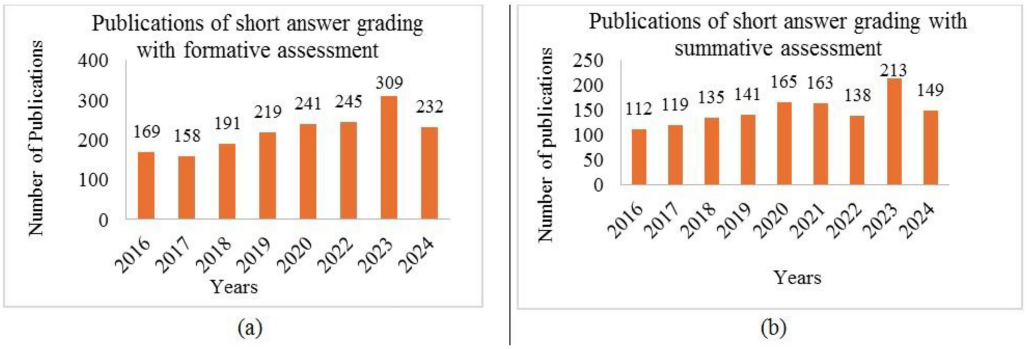

Fig. 3. Short answer grading with formative and summative assessments.

and recorded [24]. Online examinations form one of the assessment methods in an increasingly digital world, and different techniques have been recommended for automating online examination systems [9, 10, 22, 25]. Questions are administered and evaluated automatically, which reduces the time, resources and cost, e.g., staff workload or extra staffing and results waiting time [26]. Although the types of questions vary between formats and subjects, they can be almost manageable with assistance of intelligent technology. For example, **multiple-choice questions** (**MCQs**) are popular formats, and supported by systems such as Brightspace, MES, WRS, Google forms, and so on. [10, 22, 23, 25, 28, 29]. Text input methods are also available in those systems, though the function of auto assessment for text grading is missing. However, the advantages of an automatic examination system outweigh the disadvantages of using the traditional article-based methods, when considering the user experience for both staff and students [9, 10, 24, 30]. Short answering questions with text input are also available in some summative assessment systems, but the automation of assessment is still an issue to be solved. Table 2 shows a summary for formative vs. summative assessment.

Figure 3 shows a trend of research publications that report short answer grading in both formative and summative assessments. For formative assessment (Figure 3(a)), the largest increase was in 2023 for 309 articles. Although in 2024, the number is reduced to 232, the general trend has been smoothly increased from 169 to 232 for 8 years. For summative assessment (Figure 3(b)), the largest increase was also in 2023 for more than 200 articles. Likewise, the number is dropped to 149 in 2024, but the general trend has been increased smoothly from 112 to 149 for 8 years. It is clear that the publications show that ASAG systems have attracted similar attentions for both assessment formats. The evidence demonstrates that the impact of ASAG on the educational assessment systems is gradually increasing.

4 Related AI Technologies Adopted in the ASAG

A number of AI related technologies have been adapted to ASAG systems, such as Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP), Large Language Models/LLAMAs, ChatGPT, Recurrent Neural Networks (RNN), Bidirectional Encoder Representations from Transformers (BERT), and Extra-Large Bidirectional Transformer Network (XLNET).

4.1 Deep Learning Technologies with the ASAG Systems

Automatic grading systems have effectively reduced the enormous number of resources and time spent on marking tests, thus enabling teachers to optimise time and effort for other teaching duties that would result in a better academic experience for their students. Because of the exact nature of the assessment process, substantial investigations have been conducted to guarantee that a level of playing field is provided. Such a platform ought to be able to justify the replacement of human grading techniques adequately because DL has its advantages to advance ASAG systems [16, 33–35, 63].

DL architectures, such as RNNs, LSTMs, attention mechanisms, and transformer-based models, have been dominated by NLP in recent years [16, 33–35], which obtained cutting-edge results in a variety of tasks. As a result, response-based systems have gone further by utilising deep neural networks yielding significantly more promising results than other methods [32, 35, 92, 99, 140]. The results can be analysed for both formative and summative assessments [9, 10, 24, 30]. Because of using DL, current advancements in NLP may be traced back to the publications of massive pre-trained language models that are then fine-tuned for downstream analysis, a process known as transfer learning as discussed by Hossain et al [36] and Cook and Karakus [37]. Thus, DL has transformed NLP into ASAG systems [33–35]. Furthermore, the advantages of DL can overcome the limitations of traditional RNNs, e.g., the loss of collected information at the start of a sequence [141].

4.2 NLP Text Pre-processing with ASAG Systems

NLP includes the unsupervised analysis of free-text responses and enables a machine to analyse natural language autonomously. The literature demonstrates various approaches to NLP, spanning from statistics to informational formal language theory. In this aspect, **Artificial Intelligence** (**AI**) has altered the way associated tasks by utilising ML or DL approaches. ML/DL uses a variety of strategies to infer knowledge from huge amounts of data. Using such strategies enables the system to simulate the human thinking process. The ML/DL has resulted in relevant outcomes in a variety of related tasks, e.g.,

- —Translation software, i.e., converting one natural language into the other autonomously [127],
- —Evaluation of text sophistication, i.e., trying to evaluate the text complex nature of a paragraph autonomously [124],
- —Vocabulary improvement, i.e., assisting students in improving vocabulary [39],
- —Social Networking Analysis, i.e., analysis of text information from Social Media [40],
- —Independent essay scoring [141].

The use of NLP in auto grading systems shows a growing trend from 2016 to 2024 (see Figure 4). With the integration of ML, DL, LLMs and other AI-based technologies, the research outcomes were published in 14 articles in 2016 rising to 117 articles in 2024. Most users are interested in applying it for essay scoring and text processing [41–43]. As an AI related technology, NLP is popular in a number of subject areas. According to sciencedirect.com, 10 subject areas employed it, not

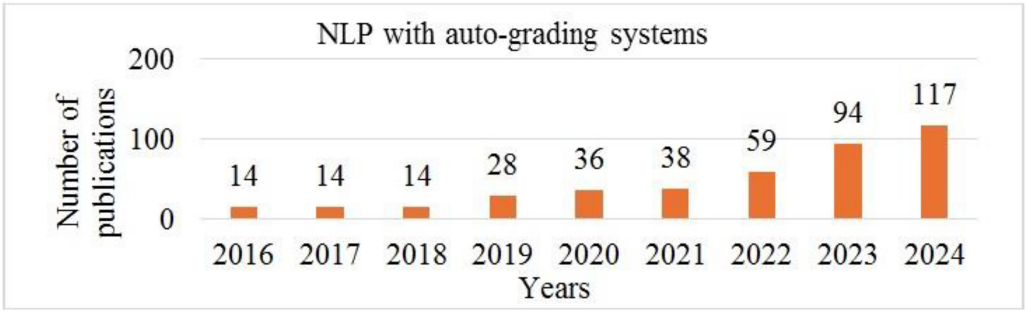

Fig. 4. NLP with auto grading systems.

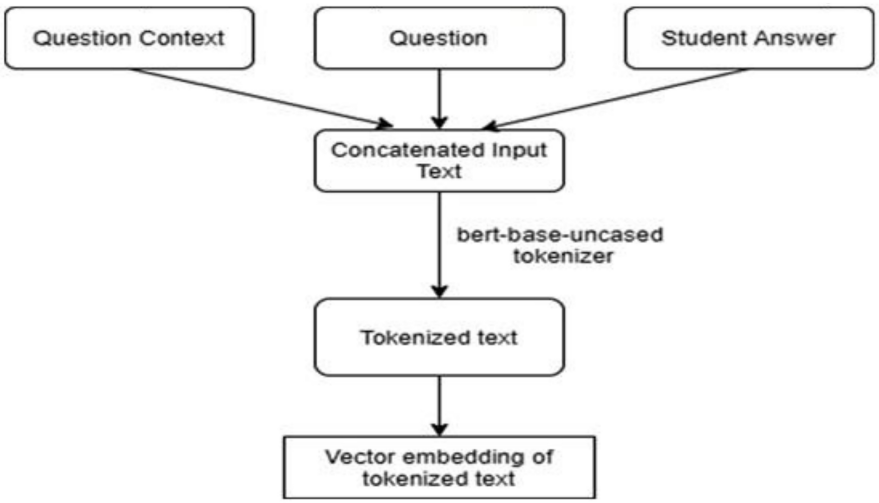

Fig. 5. Pre-processing Flowchart.

only by Computer Science, but also by engineering, medicine and Dentistry, Decision Science, social science, business, maths, and so on. The studies reported through 25 academic journals/media (sciencedirect.com accessed by 2 November 2024).

Text pre-processing is the initial step in NLP (see Figure 5). It transforms the text into a predictable and analysable format, allowing ML algorithms to perform better. Tokenisation, stop-word removal, normalisation, stemming or lemmatisation, and part-of-speech are all text preparation techniques.

Text pre-processing is the initial step in NLP (see Figure 5). It transforms the text into a predictable and analysable format, allowing ML algorithms to perform better. Text preparation techniques are tokenisation, stop-word removal, normalisation, stemming or lemmatisation, and part-of-speech.

Tokenisation and stop-word removal –Tokenisation is the process of separating or splitting a text into a token list. Tokens might be words in an expression or phrases in a paragraph. In NLP, the stop-word is a word frequently used in a language but does not contribute to the semantics of the document and has no valuable information (e.g., pronouns and prepositions). Stop words in English include “a,” “the,” “us,” and “our.” by removing low-information terms from of the text, the focus would shift to the keywords, and reduce the corpus size, resulting in greater efficiency.

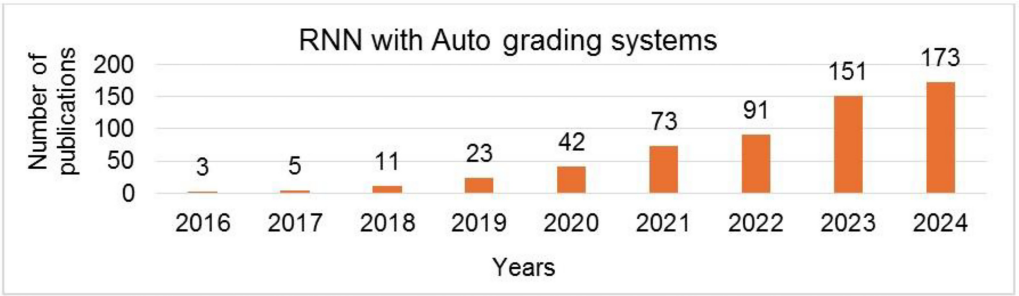

Fig. 6. RNN with Auto grading systems.

Normalisation –Normalisation is the process of transforming a text into its usual (standard) form, e.g., the terms "2moro" and "tomrw" into their normal form, "tomorrow". Another example is the mapping close words, e.g., "key-words", "key words", or "keywords" to just "keywords".

Lemmatisation or stemming – Stemming plays a crucial part in NLP when it comes to deleting suffixes or prefixes. Stemming can result in incorrect interpretation and spelling problems, but lemmatisation considers the context and changes it into its basic meaningful form, and returns a word to its basic or root form known as a lemma.

Parts-of-Speech (POS) –POS tagging refers to the process of allocating one of the speech parts to a certain term. POS tagging is a way of labelling every word inside a sentence with its correct segment of speech. This is frequently a more straightforward way for schoolchildren to identify words such as verbs, adverbs, nouns, pronouns, adjectives, conjunctions, and so on.

4.3 Recurrent Neural Network (RNN) with ASAG Systems

RNN, unlike a simple **feed-forward network** (**FNN**), remembers items from both present training and earlier inputs, which are referred to as Hidden State Vectors [44]. Depending on the prior inputs, the same input can create several outcomes, i.e., permuting the input sequence generally results in distinct outcomes. RNNs have substantially enhanced the interpretation of sequential data by accurately capturing aspects that is contained in natural language, putting into account prior words, and critically capturing the meaning in a sentence [137]. This type of network is ideal for tasks that require the assistance of a context, including speech recognition and other NLP applications.

RNN has two key features: (1) **Parameter Exchange** (**PE**), (2) **Gated Recurrent Unit** (**GRU**). For PE, the parameters share across inputs. When this type of network does not integrate them, it is simply a conventional FNN with its own weights for each input. GRU as well as the **Long Short-Term Memory** (**LSTM**) are the most widely employed in RNN [33, 72, 78].

Figure 6 shows that RNN was relatively newly used in auto grading systems, in 2016, only 3 publications are available, but in 2024 there were 173 articles published in 10 subject areas and 25 academic journals, there is an increasing trend in the auto-grading systems, according to sciencedirect.com.

4.4 Large Language Models (LLMs) with ASAG Systems

The techniques of LLMs have been used for ASAG systems in many disciplines [31, 45–47]. LLAMA-2 is the most recently reported modes that developed further based on LLMs [48, 49]. These models use a unidirectional environment and were pre-trained on large textual corpora, which give them remarkable contextual information-gathering abilities [47]. They provide a thorough comprehension of brief responses since they might represent a phrase rather than simply as

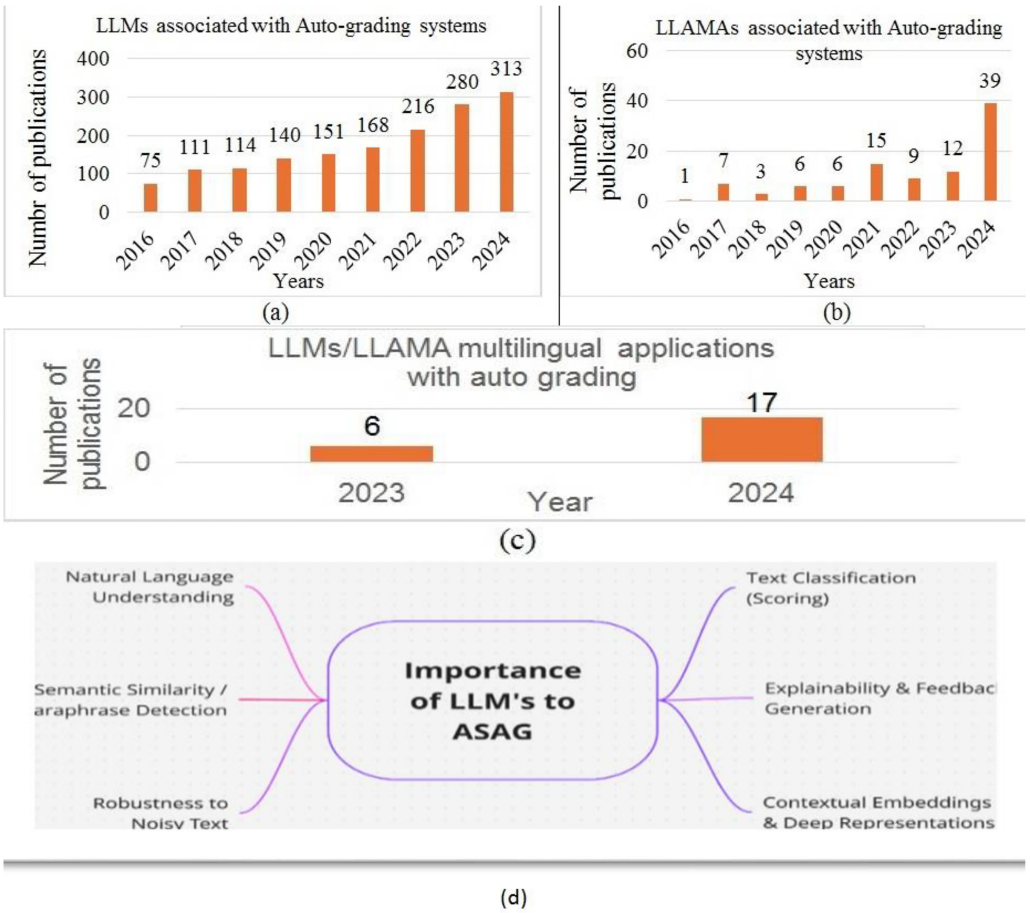

Fig. 7. LLMs and LLAMA associated with ASAG and their importance to ASAG system.

a list of words. BERT and GPT based LLMs have been optimised to produce outcomes in a variety of grading standards [46, 47, 50].

Although LLMs is a cutting age technology, the reliability of current modes is still debatable. In particular, the modes are used in the medical science [51]. Quah et al. prepared 259 questions to evaluate the accuracy of answers using LLMs together with GPT-3.5, GPT-4, Llama 2, Gemini, and Copilot [51]. They found 68.9% for the performance of 'basic science' and 45.9% for the performance of subject based questions, i.e., 'pharmacology'. They recommended that the current technologies could be used for teaching, but not advise for the decision-making in clinic practice. Of course, with a performance of 45.9% accuracy, it is not confident to make a right decision.

The technologies are getting popular according to the sciencedirect.com, about 10 subject areas adopted them and reported through 25 academic journals/media (see Figure 7(a) and (b)). LLMs/LLAMA have extended the powerful functionalities into multilingual applications with auto-grading systems (see Figure 7(c)). Correa et al attempted a small case study in the Brazilian Portuguese [143]. They concluded that the instructions from English language cannot be simply translated into other languages, thus, making additional effort is not avoidable. A recent study discussed the importance of cross-lingual analysis for political bias and false information prevalence

in a few recent tools, which are underpinned by LLMs based models or chatbots [144]. The studies involved in Russian, Ukraine and English, and evaluated with four or five LLMs based tools, i.e., ChatGPT, Bing Chat, and Bard/Gemini. They have no conclusion for which tool is performed better for using multilinguals [144].

A study investigated in the area of sentiment analysis with three languages, Czech, French, and English [145]. They used the latest LLMs based modes, e.g., LLAMA2 and ChatGPT, with positive conclusions because the results achieved are impressive at least 1% to 3% accuracy, in comparison with other existing models, such as modes based on CNN, LSTM, BERT and XLM, and so on. [145]. These technologies are developed relatively new, only 6 articles reported in 2023, 17 articles reported in 2024, according to sciencedirect.com (see Figure 7(c)).

Figure 7(d) shows the significant contributions of LLMs to ASAG, which demonstrate the **importance** of the technology that is correlated between LLMs and ASAG as discussed in recent LLM-generative approaches [41, 146–151], in particular, **for what** LLMs are doing and corresponding **to what** ASAGs are potentially needed as shown below:

- —Natural Language Understanding (LLMs) → Handle diverse student phrasing (ASAG) [146]
- —Semantic Similarity/Paraphrase Detection (LLMs)→ Match student answers semantically to reference answer (ASAG)[41]
- —Text Classification (Scoring) (LLMs)→ Assign accurate grades (ASAG) [147]
- —Explainability and Feedback Generation (LLMs) → Provide meaningful, interpretable feedback (ASAG) [148]
- —Contextual Embeddings and Deep Representations (LLMs) → Capture nuanced meaning in student responses (ASAG) [149, 150]
- —Robustness to Noisy Text (LLMs) → Deal with typos and grammar mistakes (ASAG) [151]

4.5 ChatGPT Associated with the ASAG Systems

Generative AI, such as ChatGPT, may help instructors to prepare questions and students to understand the topics [31, 52, 53]. Thus, it could be indirectly associated with outcome of assessments, e.g., formative assessment when score is involved in text based contents.

Jukiewicz reported that using ChatGPT for grading programming assignments can improve the efficiency with an automatically marking operation [54]. The article also noted that ChatGPT is impartial and unbiased. The coding standards can be enforced with the evidence presented [54]. However, Freire et al. argued that ChatGPT has an issue in reliability for generated answers, e.g., when they created 30 questions the confidence range was very low, about 22.9% and 28.6% [55]. Haman and Skolnik used ChatGPT for postgraduates to conduct literature reviews, but they argued that ChatGPT may provide many fake articles that lead to very poor results [56]. Alshehri et al. reported that in the medical domain ChatGPT may appear satisfactory to users, but generate inaccurate answers, e.g., to the questions of common patient hip arthroscopy [57]. Thus, this new technology brought challenges to users positively and negatively. Accuracy is a main concern for a number of investigations [54, 57–60]. With the intervention of generative AI, LLMs and NLP, ChatGPT are getting stronger and the popularity is increasing. It has been reported for ChatGPT, about 180.5 million users up to 2024, once reached 100 million weekly, according to SEO.AI (seo.ai, accessed 2 November 2024). Assessment with the acceptance of AI intervention will be a new challenge in future educational society.

The technology of ChatGPT is relatively new, only 42 articles published in 2023, and 124 articles published in 2024, about 10 subject areas have adopted it and reported through 25 academic journals, according to sciencedirect.com (see Figure 8).

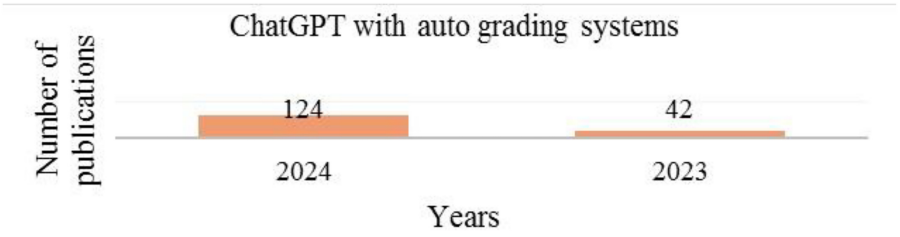

Fig. 8. ChatGPT used for ASAG systems.

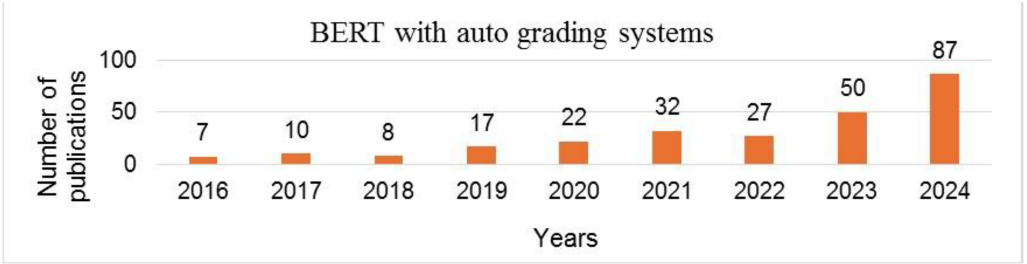

Fig. 9. BERT used in auto grading systems.

4.6 Bidirectional Encoder Representations from Transformers (BERT) with Auto Grading Systems

BERT is relatively well-known technology in text processing and has been used in auto grading systems [41, 61–63]. From 2016 to 2024, the number of publications increased from 3 articles to 87 articles. The technology has been applied for 10 subject areas and reported through 25 academic journals/media, according to sciencedirect.com (see Figure 9).

In 2019, Devlin et al. made a significant advancement in NLP by introducing BERT, which showed the value of bidirectional background in pre-trained model languages [31]. BERT is better at picking up subtleties and meanings in text because it learns situational interpretations by taking into account of left and right background [32]. New state-of-the-art standards have been established in numerous NLP assessments [31, 32]. Due to its capacity to comprehend the larger context of responses, BERT can assess responses by considering how words run collectively and individually in a cohesive response [35]. Thus, BERT performs well for the grading tasks that require a sophisticated knowledge of student responses [44, 50]. For example, the given textual data is tokenised as, Tok_2,.... Tok_m and given inputs which are encoded to be fed into the neural network to obtain the textual output T_1, T_2, and so on. The reasons for BERT's Success in Short Answer Grading Systems are as follows:

(A) Contextual comprehending: BERT is able to evaluate a student's answer based on the question's setting. It takes into account how each word interacts with all of them, which is crucial for scoring concise responses that call for a complex understanding.
(B) Gauge how semantically comparable the student's reply and the anticipated response are: BERT might be adjusted for a particular grading activity to provide a model that calculates grades depending on how the responses are similarly represented.
(C) Managing counterparts and variants: BERT is capable of managing synonyms and linguistic variants to distinguish "photosynthesis" and "the procedure of photosynthesising".

Fig. 10. XLNET with auto grading systems.

(D) Sentence-level analysis: BERT can also evaluate the cohesiveness of the overall answers by studying the relationships between sentences. This is essential for judging responses that must be rationally organised and well-structured.

4.7 Extra Large Bidirectional Transformer Network (XLNET) Associated with Auto Grading Systems

XLNET, often known as "Extra Long-Short Term Memory", is an additional potent transformer-based theory [65]. Yang et al introduced a framework set by BERT with expansion [64], called XL-NET that uses a permutation-based learning strategy in order to anticipate each word in a phrase and expands on the idea of bilateral contextual modelling [63]. XLNET can identify deeper information relationships than BERT and represent relationships between all words in a sentence [63, 64].

XLNET is ideal for short answer grading because it can capture intricate context-based linkages between words and phrases, particularly when the exam calls for students to offer a thorough justifications and background for their responses. Aurpa et al. used XLNET to reorganise mathematical equations with the results achieved 99.80% in accuracy [65]. Thus, XLNET has shown high performance in NLP tasks and it has benefits as:

(A) Enhanced environmental grasping: XLNET is resilient when evaluating the syntax of a student's answer since it can take into account of various word ordering and interdependence.
(B) Decrease in pre-training biases: The permutation-based learning used by XLNET lessens prejudices brought on by the word order in the training information. This will help with short response scoring since it enables a more equal evaluation of various word choices and phrasings.
(C) Conceptual consistency: XLNET can accurately identify phrases' conceptual coherence, guaranteeing that the responses are logically structured to make sense within the inquiry settings.
(D) Managing complicated sentences: The capacity of XLNET to capture long-range linkages is beneficial for evaluating higher-level brief responses with intricate connections.

It seems that not many articles of XLNET related to the ASAG systems are reported from the literature (see Figure 10). From 2020 to 2024, only 17 articles were published, but these small number of publications involved 7 subject areas and 15 academic journals/media, according to sciencedirect.com. The evidence showed that some potentials for the technology in near future.

5 The Computational Similarity Methods Used for the ASAG Systems

In general, computational similarity modes in this context are used to find similar text for the short answers during assessment. Ye and Manoharan studied a mode that used ML techniques with NLP

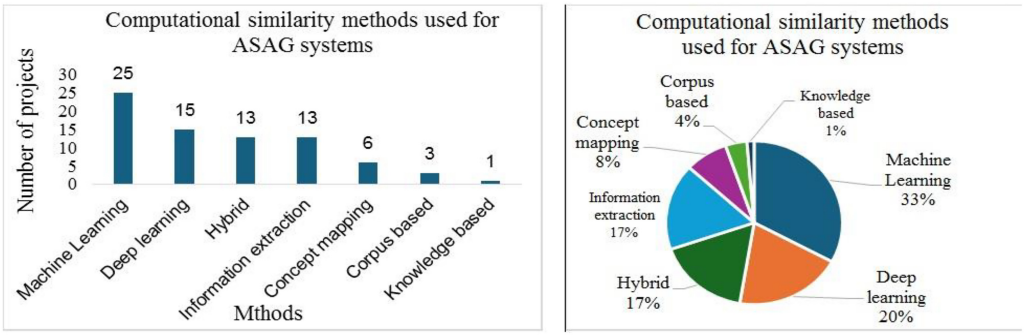

Fig. 11. Various computational methods used in ASAG projects.

to grade short answers automatically [66]. This method verifies the student answers by comparing it with a specimen answer based on the semantics presented instead of word comparison. Nael et al used a DL approach for an Arabic short answer grading system [67], which adopted a dataset based on the Arabic language called AR-ASAG developed by [68, 69]. One of the most basic and efficient strategies used to construct such systems is reference-based systems, given the student's response a reference answer [31]. Another technique employed is the similarity measurement, including cosine similarity and Levenshtein distance [32]. These measurements can then be utilised inside an algorithm to classify each answer. Farouk noted that the question and answering systems can use text similarity measures to verify the answers and assign grades [70]. Farouk further mentioned that the challenges arise while measuring the similarity and accuracy between sentences that were written in different ways [70]. Kadupitiya et al. implemented a system to assess answers automatically, according to the question types and other restrictions provided in the marking rubric [69]. Putri et al. reported that the text document feature can be extracted using Term Frequency-Inverse Document Frequency and then classified using K-Means, mainly relevant to the frequency and the semantics of words [71].

The reviewed results are shown in Figure 11. In similarity measurement, DL is used (20%) to perform the grading tasks since 2017, but ML has more users (about 33%). For the corpus-based approach, only 3 projects were involved with small ratio of users (about 4%), and used with other techniques as a hybrid approach [18, 19, 72]. For the knowledge-based approach, only one project was involved with two techniques used together, i.e., corpus and ML, as a hybrid approach [19]. About 17% projects involve the hybrid approach. Magooda et al. used more than one techniques either to improve the efficiency and accuracy or to compare the efficiency with different techniques [73]. The hybrid techniques as generic models have been used for two or more similarity measures. The purpose is to identify which similarity measurement gives better performance.

Figure 12 shows the relationship with these computational techniques. Most projects use multiple techniques. Galhardi and Brancher discussed the use of ML techniques [74]. Zhang et al. combined techniques of information extraction using ML and DL [75], whereas Marvaniya et al. used methods of concept mapping and ML [76]. Bonthu et al used ML and DL approaches [5]. A very few projects have used a single approach, such as Sreevidya and Narayanan who used DL [77]. Two projects used the combinations of three approaches: Sahu et al. used corpus-based, knowledge-based and ML techniques [19]; and Tulu et al. used corpus-based, ML and DL techniques [72]. Sahu and Bhowmick, Kumar et al used a stacked regression ensemble model that has shown improved performance over the single regression method [19, 78, 83]. Burrows et al. have done an evaluation on the methods used to grade the short answers [38]. To auto grade short answers, a deep understanding of theories in knowledge representation and information retrieval/extraction

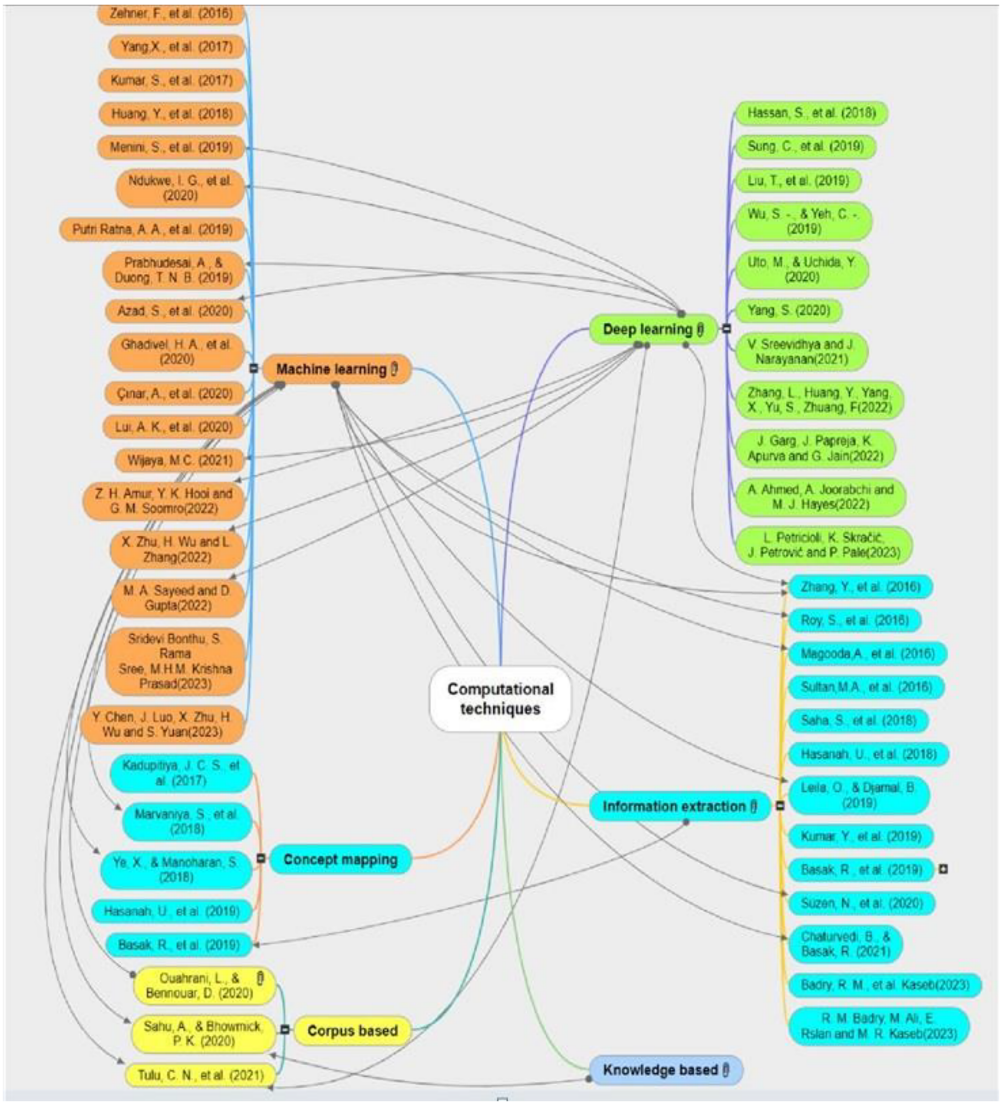

Fig. 12. The relationship with similarity modes and computation techniques used in ASAG projects.

is required to process the free text written by the students. Most modes have been developed for several years with mathematical proven (see Table 3). In most systems, feature extraction is based on the information retrieval, machine language, textual entailment, and n-grams. With technology upgrades, the hybrid techniques are suitable for the tasks.

A number of models are applied for the computational similarity [13–16, 38]. These modes are used to measure ASAG systems including word-based text similarity and structure-based sentence similarity [70, 79, 80], character based lexical similarity [13, 80, 81] and meaning based semantic similarity [80, 82], and so on., as shown in Figures 13, 14, and 15. By using the vector representations, the sentence similarity is measured (Table 3). In computation, vector-based search could avoid unnecessary paths to improve the performance as experimented by [91] that deployed the vector-based graph mode for document processing. The structure-based approach is the least used

Table 3. The Computational Modes for Measurements of Similarities and Performance

	Lexical similarity computation modes		
1.	Longest common substring similarity	L(SA, SB) = $max\ 1 \le i \le m, 1 \le j \le n Lsuff\ (S\ 11 \ldots i, S\ 21 \ldots j)$ Where,m – length of string S_A, n – length of string S_B, Lsuff – function to find the longest common suffix	[13]
2.	Jaro similarity	Dj= $\begin{cases} 0 if\ c = 0 \\ \frac{1}{3}\left(\frac{c}{\lvert s_1 \rvert} + \frac{c}{\lvert s_2 \rvert} + \frac{c-t}{c}\right) otherwise' \end{cases}$ Where, c – number of similar character, t – half of the number transposition	[13]
3.	Levenshtein Distance	S(S_A, S_B = $\frac{1-d(S_A,S_B)}{max(l(S_A),l(S_B))}$ Where, l – is the length of the string	[14]
4.	N-gram	N(S1,S2) = $\frac{n \times number of similar bigrams}{total number of bigrams}$	[84]
5.	Cosine similarity	S(a,b) = cos θ = $\frac{a.b}{\lVert a \rVert \vee b \vee}$	[15]
6.	Web Jaccard similarity	JC(y_i,y_j) = $\frac{y_i \cap y_j}{y_i U y_j}$	[85]
7.	Overlap coefficient	OC(S_A, S_B = $\frac{\lceil K \lvert S_A \rvert \cap K \lvert S_B \rvert \rvert}{min(\lvert K \lvert S_A \rvert \rvert, \lvert K \lvert S_B \rvert \rvert)}$	[86]
	Semantic similarity computation modes		
8.	Normalised Google distance similarity	GD(a,b) = $\frac{max\{logf(a), logf(b)\} - logf(a,b)}{logN - min\{logf(a), logf(b)\}}$	[80]
9.	Resnik similarity	R(c1,c2) = ln(p_{is}(c1, c2)	[97, 88]
10.	Sentence similarity measures	Word Based sentence similarity Word Matrix WM = $[a11 \cdots a1(n-\delta) \vdots \ddots \vdots a(m-\delta)1 \cdots a(m-\delta)(n-\delta)]$ $WS = \frac{\delta \sum_{i=1}^{\lvert p \rvert} pi \times (m+n)}{2mn}$	[89]
11.	Sentence similarity measures	Structure based sentence similarity Grammar based Similarity(S_A, S_B) = $\frac{2 \times depth(h_{S_A,S_B})}{dep_{pl}(S_A,h_{S_A,S_B}) + dep_{pl}(S_B,h_{S_A,S_B}) + 2 \times depth(S_A,S_B)}$	[90]
12.	Vector Based sentence similarity	Average of word $\vec{S} = \frac{1}{M} \sum_{k=1}^{M} \vec{w_k}$	[91]
	Computation performance modes		
13.	Accuracy	A = $\frac{TP+TN}{TP+FP+TN+FN}$	[66]
14.	Precision	P = $\frac{TP}{TP+FP}$	[90]
15.	Recall	R = $\frac{TP}{TP+FN}$	[17]
16.	F1- score	F1 = $\frac{2 * precision * recall}{precision + recall}$	[92]

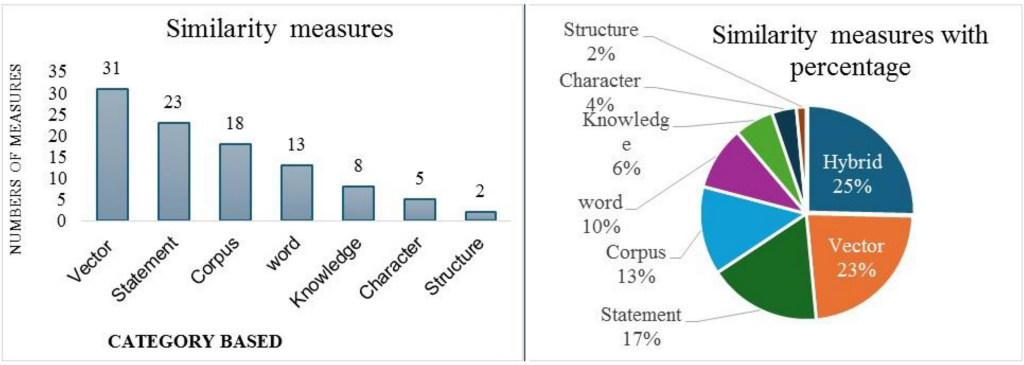

Fig. 13. The distributions of similarity measures based on different categories.

Fig. 14. The relationship with the categories of similarity measures and projects.

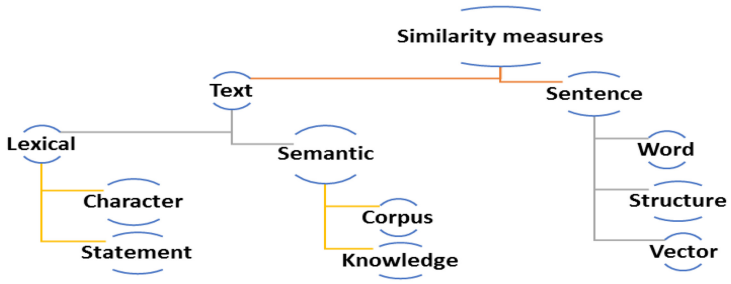

Fig. 15. ASAG systems measure different similarities.

approach and mostly considered for checking the language subjects. Meanwhile, the complexity of computation makes the approach not as popular as other methods, as evidenced by [90, 93]. Farouk and Lee et al. also realised that the challenges are involved in other elements during the computation, e.g., gamma and linguistic rules with different languages [28, 90].

Figure 13 shows the distributions of similarity measures. It was found that vector-based approaches are most commonly employed (29 projects) and structure-based approaches are the least

Table 4. The Measure Methods and Limitations Based on the Two Types of Approach, i.e., Character and Term Based

Approach	Measure methods	Limitations	Refs
Character based	Longest common substring similarity	More space used as the operation uses recursion technique.	[13]
	Jaro similarity	The result is efficient only for smaller data size.	[106]
	Levenshtein similarity	Result is not efficient for larger texts.	[15]
	N-gram similarity	This method reduces accuracy.	[107, 108]
Term-based	Cosine similarity	This method is domain dependent	[15]
	Web Jaccard similarity	Lower accuracy	[85]

reported (2 projects) in the literature. Wu and Yeh used vector-based approach by using the methods of word2vec, GloVe and sense aware vectors [17]. Figure 13 also shows that the hybrid approach is the most used (34 projects) for 25%, and the statement-based (23 projects) for 17%, vector-based (31 projects) for 23% and corpus-based (18 projects) for 13%. The measures are not frequently used in structure-based (2 projects) for 2%, character-based (5 projects) for 4%, knowledge-based (8 projects) for 6%.

Figure 14 shows the relationship with similarity categories**.** Investigations in text similarity are more popular than the investigations in sentence similarity. Sultan et al. investigated text similarity for lexical and character-based approaches and vector-based approaches for sentence similarity [95]. Tulu et al investigated statement-based, word-based and vector-based approaches for sentence similarity [72], whereas Uto and Uchida only investigated word-based and vector-based approaches [96]. Menini et al investigated three approaches, i.e., corpus-based to measure semantic similarity, word-based and vector-based approaches to measure the sentence similarity [97]. Most investigations used multiple approaches to measure both text and sentence similarities, which indicate the complexity of computational processes in comparison with character-based approach only [72, 95–97]. The researchers who used DL methods have tried to adopt at least one approach in measuring text-based similarity and vector-based in sentence similarity. Azad et al. investigated character-based approach in text similarity and vector-based in measuring sentence similarity [98]. Zhang et al. utilised statement-based, corpus-based for measuring text similarity and word-based for measuring sentence similarity [99]. A few researchers adopted just one approach for measuring similarity [19, 62, 77, 100–102], e.g., only used vector-based to measure sentence similarity; only adopted statement-based approach [34]; only used corpus-based for measuring semantic similarity [63, 103]; only used knowledge-based approach for measuring semantic similarity [104, 106]. Very few investigations adopted a structure-based approach, and character-based approach to measure the text similarity [95, 98]. This further explains that the limitations identified in the Tables 4 and 5 could be the challenges for researchers to face in near future.

Figure 15 summarises a general structure of the similarities in ASAG systems, i.e., word-based text similarity and structure-based sentence similarity [70, 79, 80], character based lexical similarity and meaning based semantic similarity [13, 80, 82]. The similarity measures can be grouped by text and sentence. The text similarity is linked to lexical and semantic based. Semantic similarity is corpus and knowledge based. Sentence similarity is word, structure and vector based.

Table 3 presents 16 computational modes that are developed or adopted. The modes are relevant to the computations of lexical similarity (1 to 7) and semantic similarity (8 to 12), as well as performance computation (13 to 16). These modes define the features of similarities in terms of strings, sentences, meanings, and words. There are four modes paid special attention to the accuracy, precision, recall and scoring with the consideration of true and false positives. From the

Table 5. The Algorithms used to Measure Semantic Similarity based on the Corpus and Knowledge Approaches

Approach	Algorithms	Limitations	Ref.
Corpus-based	Latent semantic analysis similarity	The computation is invisible	[109]
	Normalised Google distance similarity	The results are unstable if there are a greater number of Google pages	[29, 80]
Knowledge-based	Resnik similarity	Values are limited to corpus. Semantic meaning is not presented properly.	[87, 80, 110]

review of similarity measures the articles were chosen based on the analysis of short (few words), medium (up to 2 sentences) and long sentences (paragraph or essay), the domains of subjects and languages to show that the similarity measure is essential in short answer grading.

Table 4 presents the methods and limitations based on the two types of approach, i.e., character-based and term-based. The limitations are mainly related to the efficiency for large texts, extra unnecessary space used lower accuracy, and domain dependency, according to the choice of measures. There are four methods reported in character-based approach (see Table 4). Majumder et al investigated longest common substring similarity, but they found that the method needs to use more space because of recursion process [13]. Vijaymeena and Kavitha reported their approach of Jaro similarity that was efficient but only for smaller data size [106]. Olowolayemo et al used Levenshtein similarity and found that it is not efficient for large text [15]. Potthast et al and Franco-Salvador et al reported that they used the method of N-gram similarity with a reduction of accuracy [107, 108]. There are two methods reported in term-based approach, i.e., Cosine similarity and Web Jaccard (see Table 4). Olowolayemo et al investigated a method of Cosine similarity and identified that the method is domain dependent [15]. Chung et al used a method of Web Jaccard similarity and found that the accuracy was lower [85].

Table 5 presents different algorithms used to measure semantic similarity based on the two approaches, i.e., corpus and knowledge. The limitations cover (1) computational visibility, e.g., Nau et al investigated the algorithms for measuring corpus-based similarity of Latent semantic analysis [109]; and (2) stability of results, e.g., Pradhan et al used Normalised Google distance similarity and found that the values are only based on the corpus, e.g., historically reported from Resnik [80, 87]. The algorithms of Resnik similarity were further investigated by [80, 87, 110]. These reported limitations imply that the semantic similarity measures are still at very early stage, e.g., only studied at the research lab. Development of robust and reliable algorithms could be a future research challenging task.

6 Grading Methods Associated with ASAG Systems

Grading is one of the most essential components of evaluating the short answers. Grades provide the performance outcome of students' answers. All student answers vary as they write the answer in their own understanding. Hence, the teachers need extra effort to evaluate the answers carefully to provide fair grading. This is time consuming and heavy workload for the markers [26]. Automated grading was introduced to reduce the teachers' workload by using various methods [15, 17, 38, 63, 76] (see Table 6). Wu and Yeh used correct-incorrect, i.e., a 2-way system, to achieve an accuracy of 91% [17]. Marvaniya et al. reported that a 2-way approach achieved the highest figure for the classification of macro-averaged features in comparison with 3 and 5-way approaches [76]. Burrows et al. discussed how the system varies as 2-way grading, 3-way grading and 5-way grading [38]. Several systems used point-based grading as 0-1 or 0-5, and hybrid systems, depending

Table 6. Definitions for Grading Systems

No	Grading methods	Description
1.	2-way grading	Correct and Incorrect
2.	3-way grading	Correct, Partially Correct and Incorrect
3.	5-way grading	Correct, Partially Correct, Contradictory, Irrelevant and Non-domain
4.	Points based	0-5 (Mostly 0 is considered as low score and 5 is the maximum score, sometimes vice versa)
5.	Hybrid	With more than one of the above methods

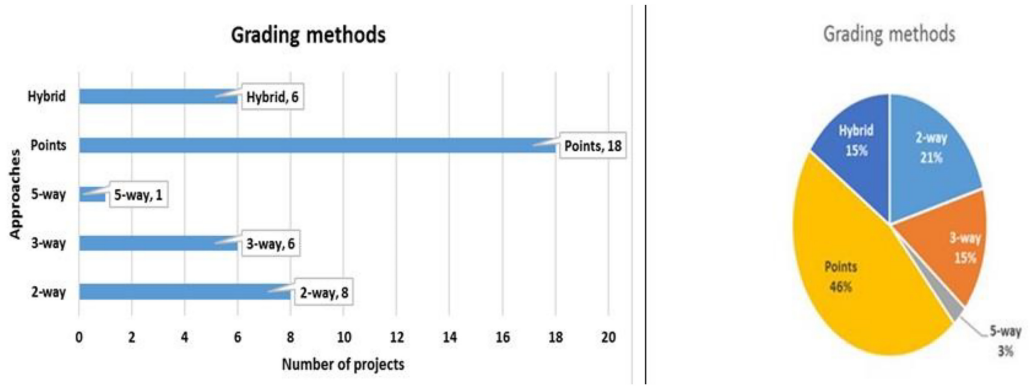

Fig. 16. Grading methods and usage percentage.

on the grading approach [66, 111–114]. Table 6 summarises the various methods. Figure 17 the grey line connections represent incorporation of hybrid grading methods.

Generally, once the similarity is measured, based on the output of the similarity level the grading is assigned. The results for five commonly used grading methods are presented in Figure 16. The most used method is scoring/points (about 46%) for 34 projects. Only one project used a 5-way method (about 3%), but 5 projects merged 5-way with other methods to improve the performance. 6 projects used more than one grading methods, i.e., a hybrid method, to evaluate the efficiency of the system [76, 115]. The purpose of grading is to record the performance of teaching and learning for both teachers and students. This section has implied a pedagogical significance, i.e., the outcome of summative assessment and formative assessment. With the point-based assessment, the collected points can be used as the records of student performance or the feedback of teaching delivery as well as the indication/guidance/motivation of self-regulated learning, especially with quiz type that is widely used in the training or online sessions [9–11, 21, 139].

Figure 17 shows the relationship with grading methods and reviewed investigations. It seems that point-based grading is the most used approach, but it is debatable because it is very much depending on the requirements of users. If other methods are not mature, it is an easy option. 5-way is the least used approach [78]. Çinar et al, Süzen et al and Wijaya reported their experience in using the point-based grading [8, 116, 117]. Marvaniya et al, Saha et al used three approaches, i.e., 2-way, 3-way and 5-way [76, 115]. Kumar et al used 5-way approach [78]. [16, 44, 69, 75] only used 3-way approach.

7 Datasets Used for ASAG Systems

Various datasets have been used in the research articles. A large number of researchers have chosen pilot datasets in their investigations to prove the system efficiency toward real-world datasets,

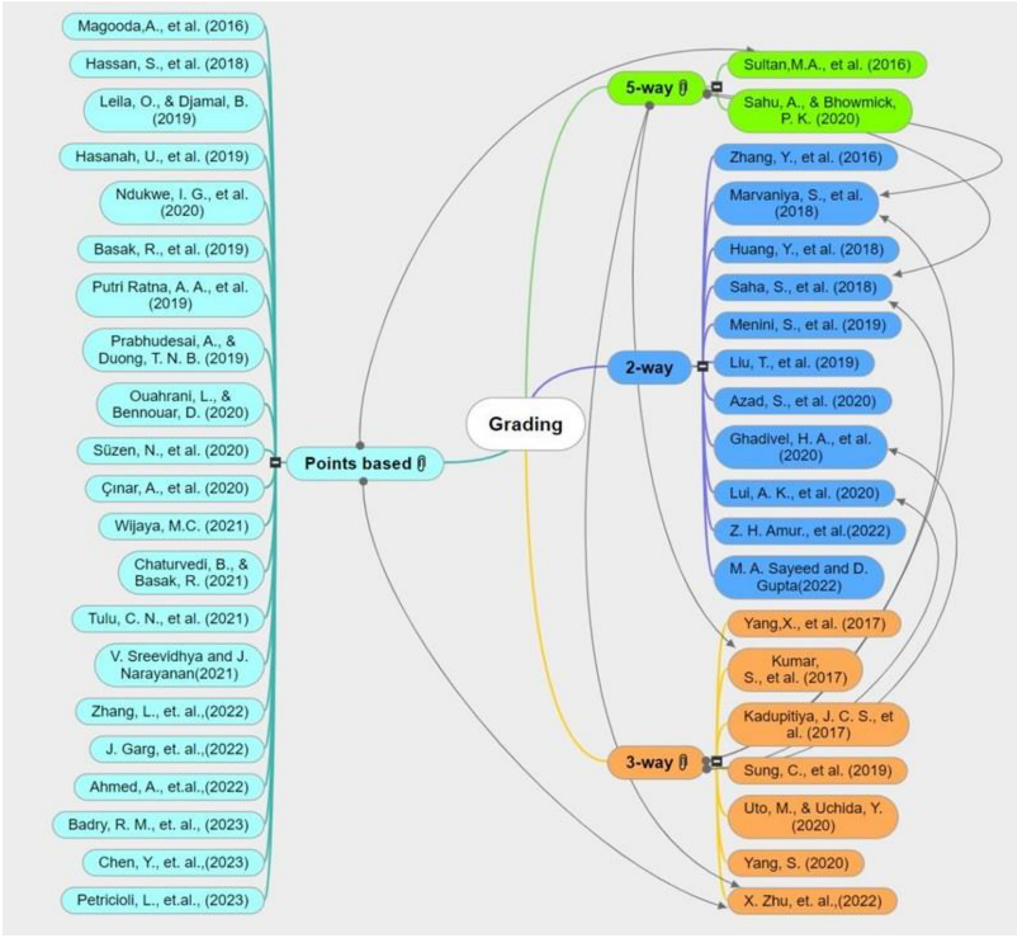

Fig. 17. The relationship with grading methods and reviewed.

because the pilot data are collected from real-world and the other named datasets are the publicly available datasets [111, 112, 118–121]. Some datasets have been developed and made available for other researchers around the world, such as, the datasets developed by Mohler et al. [122], including Kaggle [123], which has been used by several researchers [33, 78, 83]. SciEntBank has been another choice for investigations [17, 19, 63]. Recently, Sung et al. [100] used a large-scale industry dataset consisting of three domains: (1) Physiology of Behavior (Phy), (2) American Government (Gov), (3) Psychology – Human Development (Psy-I), and Abnormal Psychology (Psy-II).

Figure 18 summarises the various datasets are used in the literature. 28 approaches about 54% used pilot datasets. It seems that the pilot datasets form the most popular approaches. About 7 approaches used SemEval datasets. 6 approaches used Mohler dataset [122]. 2 approaches used the Cairo University dataset [68]. 2 approaches used Kaggle datasets [123]. 5 approaches used project individual datasets [16–19]. 1 approach used large scale industrial dataset [124]. Datasets are available not just in English, but also in other languages, e.g., the AR-ASAG Arabic dataset [67].

In general, a good ASAG system needs good quality of datasets to achieve reliable results to measure its efficiency and accuracy. It is not just about the techniques, e.g., various NLP, ML, and DL modes used to assess short answer grading, ultimately, the pedagogical quality assurance

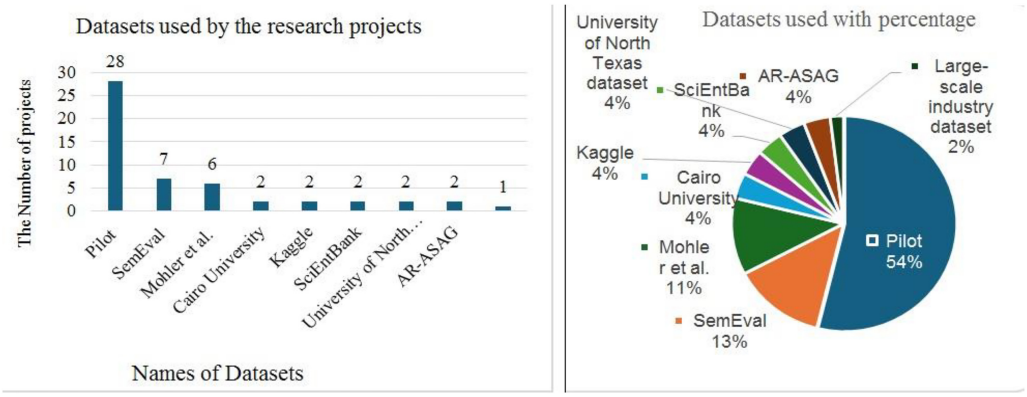

Fig. 18. Datasets used with numbers and percentages.

needs it. There are several datasets available for public use, such as the Texas computer science dataset [125], the Extended Texas computer science dataset [122], the Cairo University dataset [82], the SemEval 2013 dataset [101], and so on. Meanwhile, the advancement of technology in computer science continues keeping up the interest toward the potential enhancement of short answer grading.

8 Evaluation Methods Associated with ASAG Systems

The evolution of technology has kept the topic of ASAG still open for research. Williamson et al. argued that a major concern is trust in the system and the accuracy of the grading [113]. For ASAG, evaluation is normally a necessary process for almost every research project during or at the final stage of the investigation, particularly focusing on the observation of efficiency and accuracy [68, 76, 95]. Magooda et al. used more than one technique either to improve the efficiency and accuracy or to compare the efficiency of different techniques [73]. Burrows et al. provided a historical evaluation on the methods used by researchers to grade the short answers [38]. It is noticeable that the hybrid techniques as generic models have been used for two or more similarity measures to identify which similarity measure gives better performance. The Pearson Correlation method has been used for the model of information extraction [95], the sentence embedding techniques [76], the corpus based semantic approach [68], and unsupervised vector space approach [18]. Some attention has been paid to the evaluation of accuracy through comparison between human grading and automatic evaluation, as well as student feedback using Likert scales, and so on. [15, 44, 66, 71, 98, 126, 127].

Figure 19 summarised detailed evaluation methods and corresponding results reported from 42 articles involved multiple approaches e.g., some projects used more than one method (see Figure 20). Evaluating accuracy forms the most focus in the reviewed research articles. Figure 20 shows about 22 evaluation methods reported in the literature. One project may use multiple methods to evaluate their system. The colour lines are linked to the methods during the evaluation process, Roy et al. reported that they used methods of Skyline and Baseline [128, 129], Sayeed and Gupta used accuracy, Macro-average F1 and Weighted-F1 [130], and Ahmed et al. used Pearson Correlation in addition to RMSE [132].

Figures 21 and 22 show further cases with the accuracies attained from these studies. We reviewed 37 evaluation methods that have been analysed and their limitations are identified, which cover a wide range of problems from project to project, e.g., limited in one language and not tested in other languages [68]; domain dependent [100]; only applicable for a small dataset

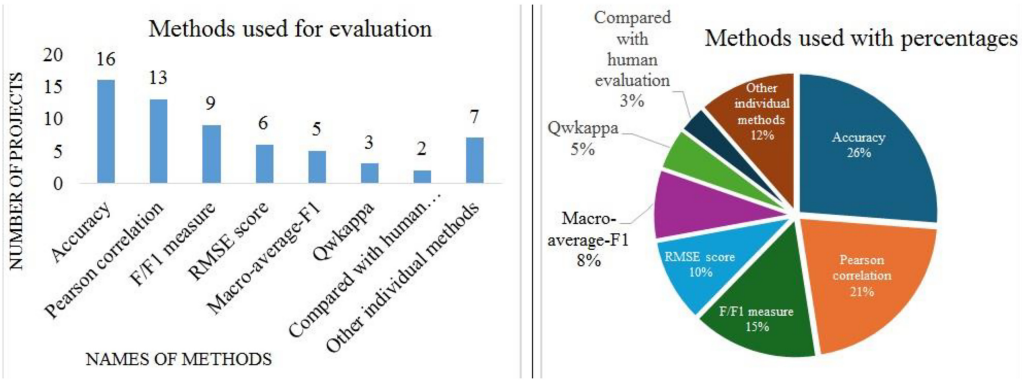

Fig. 19. Evaluation methods used and their usages by percentage.

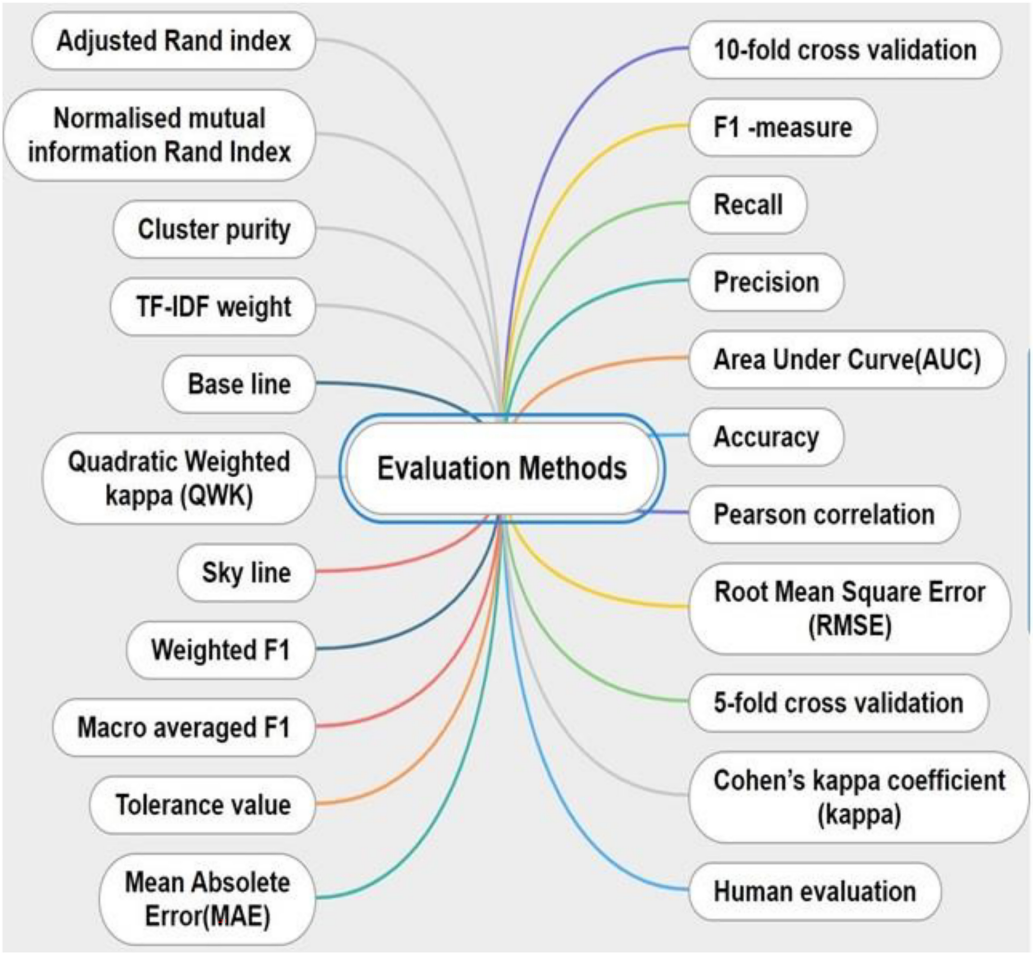

Fig. 20. Evaluation methods involved in ASAG.

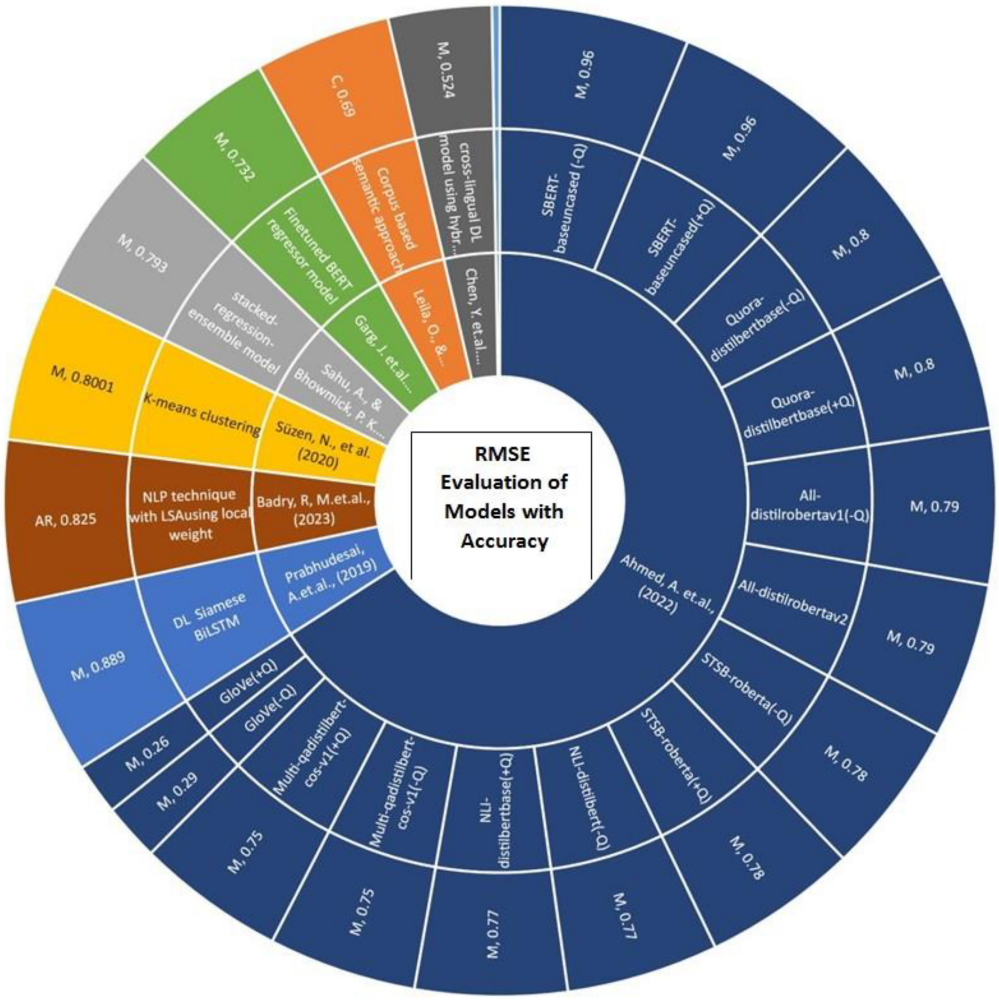

Fig. 21. The relationship with evaluation mode RMSE and accuracy.

[15]; efficiency and accuracy issues [17, 86], technology readiness issues [98], and so on. The information collected could be a useful reference to researchers in the topic area when they are going to decide which evaluation method will be suitable to their application.

Figure 21 shows the relationship with RMSE Evaluation models and accuracy, where the RMSE stands for Root Mean Square Error. Figure 22 shows the relationship with **Pearson Correlation** (**PC**) mode and accuracy. In these figures, the inner layer is the Reference; middle layer is the models, the outer layer is the evaluation score represented with datasets, i.e., M – Mohler's dataset [122], CC – Chinese Computer Network ASAG dataset [16, 17], C – Cairo University dataset [82], SE – SemEval Dataset [101], and SE-AR – SemEval Arabic dataset [14, 18, 67].

Ahmed et al reported that they have attained the RMSE score as 96% of accuracy and the Pearson Correlation 59% of accuracy [132]. As the same approach, Chen et al attained 55% of using RMSE and 97% of using Pearson correlation for their Cross-lingual deep learning using hybrid neural network model [133], in which the Pearson correlation wins by providing the maximum correlation between the answers. Both models have used the same Mohler's 2011 dataset [122]. Thus,

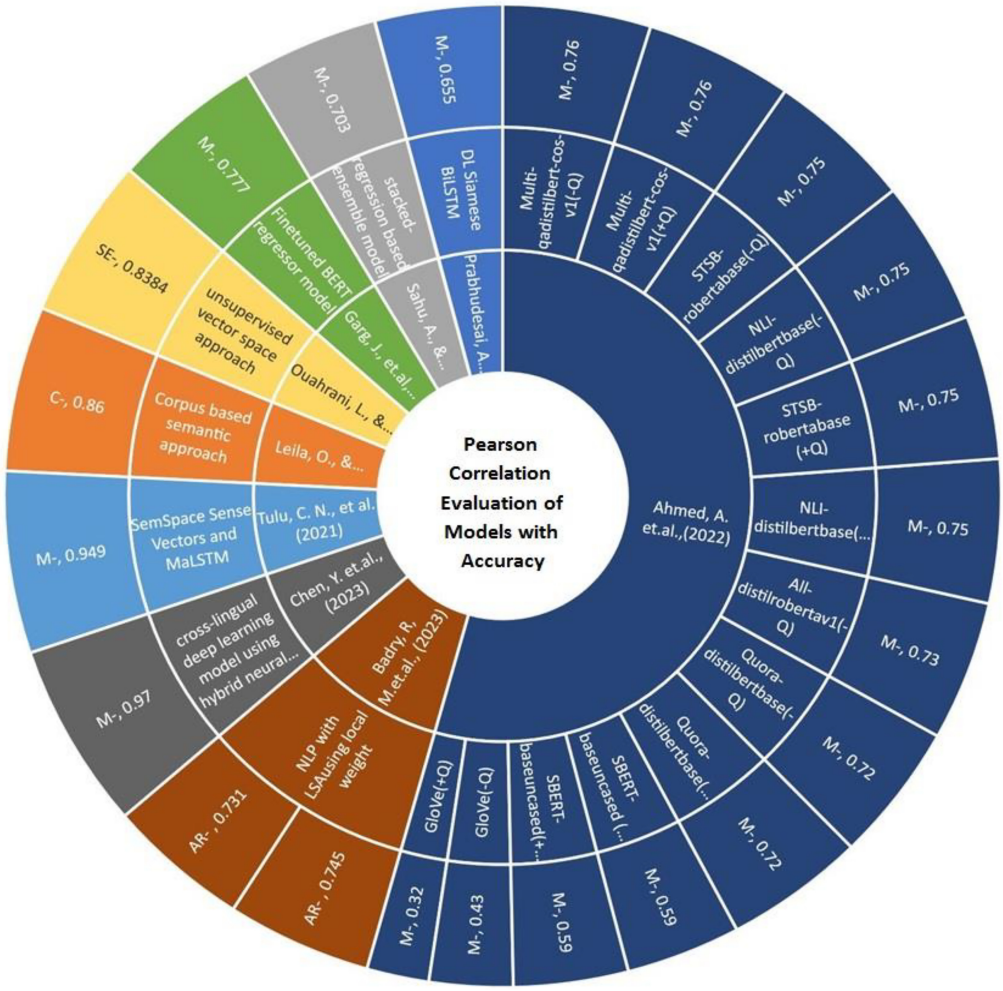

Fig. 22. The relationship with evaluation mode RMSE and accuracy.

we could say that the model needs to be evaluated using various evaluation methods to bring a conclusion on the computational accuracy of the model. Determining the efficiency of the ASAG systems is complicated with the involvement of multiple grading schemes, multiple datasets and multiple evaluation techniques [19]. While considering the grading systems, the 5-way grading system could be better than the 2-way grading system as the students could be scored for partial marks for their answers and at the same time the accuracy of grading partial marks should be evaluated. To score the final mark, there is no standard solution, e.g., a few systems used teachers' answers as the only reference answers, whereas others just used the teachers' answers as a guide for grading the answers. Thus, there are plenty of rooms to research into this area to develop an objective grading system with pedagogical theory proven.

Figure 22 is organised as: the inner layer is the Reference; middle layer is the models, the outer layer is the evaluation score represented with datasets, i.e., M – Mohler's dataset [122], CC – Chinese Computer Network ASAG dataset [6, 17], C – Cairo University dataset [82], SE – SemEval Dataset [101], and SE-AR – SemEval Arabic dataset [14, 18, 67].

Summary

Initially, 4,084 articles were identified using the keyword search. After filtering, the number of articles was reduced to 57. Based on the questions set up in Section 2, we have reviewed:

(1) A number of existing technologies associated with ASAG.
(2) AI technologies and computational methods/modes are used in ASAG.
(3) Published algorithms and their limitations.
(4) Grading methods with the definitions in Table 6, using 2-way, 3-way, 5-way, hybrid methods.
(5) Datasets and relevant issues used in the ASAG projects.
(6) Evaluation methods for ASAG systems.

The results in Figures 2 to 22 and Tables 3 to 6 demonstrate that there are 7 AI related technologies, 16 computational models of similarity measures, 5 grading methods, 7 types of datasets, and 22 types of evaluation methods with the accuracy and efficiency.

9 Discussion

From the reviewed literature, we identified that ASAG does contribute to both formative and summative assessments, evidenced by Table 2 and Figure 3(a) and (b) (see Section 3). The trend is smoothly growing and the significant benefit is to improve:

—Classroom learning and self-regulated learning – formative assessment.
—Anti-bias marking, the reduction of staff workload and results waiting time – summative assessment.

To target at the **RQ1**, we reviewed technologies that are associated with ASAG, especially recent developed generative AI – LLMs/LLAMA series, e.g., LLAMA-2, LLAMA-4, ChatGPT-4 (Section 4) and their importance and potential needs to ASAG, shown in Figure 7(d), discussed in Section 4.4. We have also looked at other technologies that are adapted into auto grading systems with case studies, such as BERT, XLNET, RNN, and relatively well employed ML, DL neurone network, evidenced by Figures 4–10. Although the new technologies mentioned above are still at the research stage, these cutting-age technologies demonstrate a growing trend for future ASAG systems that will ensure the efficiency of teaching and learning, as well as knowledge delivery toward a better standard in pedagogical circles.

To answer **RQ2**, we focus on the similarity measures evidenced by the Table 3, and Figures 11 to 15. Technically, ML / DL are most used methods to achieve the better performance, whereas corpus and knowledge-based techniques are less popular methods. Based on the literature, we compared and analysed the limitations in character-based, term-based, corpus-based and knowledge-based approaches, evidenced by Tables 4–5. We found that there are several challenges in the current available systems, e.g., (1) the lower accuracy for assessed results, (2) domain dependent for applications, (3) only applied for the small datasets, and (4) limited datasets available, and so on. Therefore, improving the technology reliability is an immediate task for computational scientists to enhance the confidence for the end users, which will be directly associated with both summative and formative assessments.

In the reviewed projects (see Figures 13 and 14), the statement-based approach is the second most commonly used approach (about 23%), and the corpus-based approach is in the third place (about 18%), with the vector-based approach being the largest used approach (about 31%). The structure-based approach is the least used approach, representing only about 2% of the research projects. Tables 4 and 5 indicate the technical limitations of the existing algorithms and approaches

for similarity measurements, which are directly relevant to the pedagogical practice as listed in Table 2 of Section 3.

We found that although the existing systems have attained a level of accuracy, each system has its own issues to research into the area before releasing to the real-world. Therefore, there is no generalised system that could be used at schools for online examinations. For instance, with the recent Covid-19 situation, the UK GCSE examinations in 2021 had been cancelled and the grading for the students was going to be analysed using teacher assessed grading, according to BBC News (2022) [134]. Thus, there are several controversies going on as teacher assessment may not be objective, e.g., favouring certain students [15]. As a result, auto-grading is a way to remove this biased assessment caused by human intervention. Another issue is found that vector-based similarity is used for analysing the semantic meaning rather than the syntactic similarity of the sentences (see Figures 13–15). Using semantic analysis, grading the syntactic method would not be very efficient as the student's answers would be in their own words based on their understanding. Hence, the semantic similarity to analyse the answer and evaluate the grading should fit the purpose to be better reflecting students' learning. The gaps and challenges we identified for the similarity measures and auto-grading performance are:

- Lack of exploration of structure-based techniques: Figures 13 and 14 showed a small number (2%) of projects reported. It raises the issue of whether structure-based methodologies for short response grading should be used if the technology is not ready for a real educational setting.
- Limited knowledge-based approach: According to the literature (only 6% projects reported), an approach based on knowledge has only been employed in one study and coupled with other methodologies. Evidence in Figures 13 and 14 concerns the possible benefits of domain-specific approaches and rule-based platforms to enhance the efficacy of automated grading.
- Inadequate investigation of hybrid methods: Although hybrid methods are said to be the most often utilised technique, there hasn't been any in-depth investigation into the precise combinations and missing information for how much they improve grading accuracy. To fully grasp the intricacies of hybrid techniques, including the precise mixtures to observe the real impact on the auto-grading performance, further study is required.

The grading method is relevant to answer the **RQ3.** We compared five commonly used methods (see Table 6). It was found that the point-based grading methods are the most used (about 46%) (see Figures 16 and 17). The 5-way approach only is the least used method (about 3%) [135], but for other projects, though they used a 5-way method, it was merged with other methods as a hybrid approach. Currently, the hybrid method is as popular as the 3-way method (both about 15%). The grading strategies used by the researchers are varied as 2-way, 3-way, and 5-way methods. The 2-way method gives whether the answer is correct or incorrect, whereas the 5-way method gives an in-depth analysis as whether the answer is correct, partially correct, incorrect, non-domain, or irrelevant. The point-based grading method is defined in Table 6. The brief response grading system's operation across a variety of models and train situations have been thoroughly covered by [138]. The models are trained using a dataset created especially for scoring brief answers, which helps to adept at the subtleties of this kind of work. It follows that the following challenges have been identified:

(1) Insufficient variety in assessment methods: According to the research, scoring/points method is the most often used grading system, with little investigation of alternatives evidenced in Table 5, Figures 16 and 17. This raises concerns regarding the variety of

grading techniques and how well they may be used in various educational settings and topic areas. Thus, grading type is another challenge as currently the reliability of evaluation is debatable, i.e., which type could reasonably reflect the students' learning outcomes is not known. A recurrent network-based approach that simultaneously learns query and response interpretations was suggested [135]. They highlighted the significance of DL in grading systems by demonstrating a superior performance over the conventional rule-based techniques. Furthermore, in order to attain cutting-edge outcomes, Liu et al underlined the possibility of fine-tuning transformer representations as BERT and GPT based LLMs for grading system [35]. The introduction of transformer-based designs has had a significant impact on how Question-Answer Generation – QAG networks have evolved [142]. Transformer-based models' introduction has signalled a paradigm change in grading system [100]. These models have shown to be capable of deciphering subtleties and context in assessment, which has enhanced grading precision. Thus, achieving cutting-edge outcomes has relied heavily on fine-tuning and transferring knowledge.

(2) A reliable dataset is important for grading performance. For the datasets selected from the reviewed literature, we found that the pilot datasets are the most applied, e.g., 42 (about 54%) projects, more than half the projects used them (Figure 18). It was also found that only one project used a large-scale industrial dataset. That dataset consisted of three domains: (1) Physiology of Behaviour (Phy); (2) American Government (Gov); (3) Psychology – Human Development (Psy-I) and Abnormal Psychology (Psy-II), which was specifically for the purpose of conducted project. The datasets also revealed a practical issue of implementing and evaluating ASAG systems, i.e., if the datasets are not available in other languages, the system can only use a pilot dataset for a particular language. While considering the linguistic systems, for the diacritic words, semantic analysis is less accurate. This is because if the language does not have a proper corpus, the analysis of the semantic meaning with various words becomes complicated. Hence, the researchers of the linguistic systems must be aware of the above discussions.

The following discussions are relevant to **RQ4**. Evaluation metrics are used to evaluate a model's performance, including accuracy and other pertinent indicators. Thorough examination of the outcomes illuminates the advantages and possible drawbacks of the suggested methodology. It is clear that most researchers have developed their systems and demonstrated the efficiency, but the systems have inherited their own disadvantages, e.g., the methods are not generically evaluable, language dependent [68, 82], or domain dependent [100]. To guarantee the efficacy and conformity to human evaluation requirements, the systems must be continuously evaluated and improved via progressive feedback cycles involving teachers, users or specialists and educational practitioners.

Pedagogically, the most critical issues raised are mainly related to the assessment accuracy that brings a huge challenge into the debate. The accuracy will directly impact on the maintaining a QAA required good quality/standard of assessment and reflect the real learning outcomes for both ability and knowledge in the subject areas. Although in recent years, several cutting-age technologies (see Figures 4 to 10), such as LLMs/LLAMA-2, ChatGPT-4, have been introduced into ASAG systems, the results for assessment accuracy are not satisfactory, as discussed in Section 4. Thus, some users recommended that the applications are not ready for the assessment, though the technologies are advanced, especially in the medical science subjects, e.g., as the performance achieved is only 45.9% that is far below the required confidence to make decision as reported by [51]. Further evidence demonstrated in Figures 21 and 22 did show that the measurement of accuracy is the most considered component during the evaluation process. However, the RSME method has achieved 96% accuracy, but it is not as popular as Pearson Correlation – PC method

that is achieved about 76% accuracy reported by [132]. From the review on online examination systems, not many projects have reported on the short answer grading. To achieve better accuracy, more clarifications are needed to process short answer sentences with conjunctions, connectives, collocation, and presence of indistinctness. In fact, not many systems have been explored by the teachers, which might be the reasons or barriers to obtain the satisfactory accuracy in the pedagogical point of views. It follows that automated grading is not a simple assessment process. An accurate formative assessment may significantly benefit learners to receive the effective learning guidance, e.g., Self-regulated learning [9–11, 21–23]. An accurate summative assessment is critical to decide the learners' pass rate and the final image to the teachers for the successful knowledge delivery. With the important ramifications for contemporary learning, it will take pedagogy, fairness, and technical factors into account. By addressing these issues, strong, effective and reliable computerised systems need to create that can help students and teachers in academic practices. It follows that pedagogically, ASAG systems will directly influence the areas of (1) both formative and summative assessment; (2) effective learning and teaching; (3) adequate feedback to the learners; (4) associated to the designed learning outcomes in the educational institutes [4, 9–11, 22–24, 28, 30, 97].

In the light of above, to answer **RQ5**, the following recommendations are proposed:

(1) Improve the level of technology readiness for the existing and newly published advanced systems to establish stronger confidence to increase users' acceptance from cross-sectors, cross languages and cross board applications.
(2) Improve the quality of similarity measures, especially, fill the gaps of structural, knowledge and hybrid approached to enhance the functionality of ASAG and quality of assessment.
(3) Increase the reliability and variety of grading methods to achieve the objective assessment as much as possible. Currently, point grading is limited to be good at summative results, but a feedback-based approach is helpful to learners in learning guidance, e.g., SRL.
(4) Establish quality datasets to be covered for a wide range of users who are from different disciplines, educational backgrounds and geographical locations to increase the credibility of analysed outcomes. It is better to be accredited by the professional bodies rather than randomly reporting with limited generic significance.
(5) Enhance the evaluation of computational analysis and grading methods to maintain a good standard of assessment accuracy / efficiency and to speed up deploying newly developed advanced ASAG tools and technologies into next generation's educational settings.

10 Conclusions and Future Considerations

A detailed study on ASAG was carried out reflecting its challenges, strengths, and limitations. After the systematic review, we identified the limitations of existing models and algorithms evidenced by Tables 3 and 4. These identified limitations are research gaps as discussed in Section 9, which will be the useful references to benefit academic researchers, pedagogical practitioners and industrial colleagues who are interested in developing a solid ASAG model to fill the gap or using the applications. The accuracy of auto-grading systems will impact on the summative assessment and formative assessment that is associated with the feedback to the learners, designed learning outcomes, as well as the measures of effective teaching and learning in the educational institutes. In general, integrating AI technology into educational system is inevitable. Likewise, grading systems are going to catch up the trend from modelling, analysis to application development in the near future. Another point needs to be addressed at the final remarks, i.e., hybrid approach is a trend in improving the quality of grading similarity measures, accuracy analysis, and efficiency of

Fig. 23. Automated grading system.

evaluation process. From the evidence presented in sections, numbers of investigations are looking for the solutions from multiple approaches; either it is computational modelling or an evaluation process.

To answer the research questions (**RQs)** proposed at the beginning of the article, we identified:

(1) The most popular techniques (**RQ1**), i.e. AI based techniques: ML/DL, which is a trend in the interdisciplinary research, e.g., computer science and educational society.
(2) The theoretic modes of similarity measures (**RQ2**), which provide a foundation or guidance for computational modelling in the topic areas, evidenced in Table 3.
(3) Commonly used grading methods and tools (**RQ3**), although some tools are arguable, e.g., point based grading, it is the most popularly used in the reality because other alternative tools are not ready to the markets, evidenced by Table 6 and Figures 16 and 17. It implies that there is an immediate need to develop ASAG systems including other types, which are easy and ready to use with better accuracy in educational circle. The more choices of datasets could be better suitable for the educational purposes, because the limited datasets currently available in public domain are inadequate to the real situation. Thus, most users select to use pilot datasets with their own format and structures, evidenced in Figure 18, which bring a new challenge to computational modelling.
(4) Most existing challenges in evaluating ASAG aim at improving the quality of feedback of formative assessment and accuracy of summative assessment in the view of pedagogy (**RQ4**). A good quality of feedback can be an effective guidance for learners to improve their SRL and teachers to improve their delivery strategies. Meanwhile, in higher education, accurate summative assessment is critical to the final decision of learners' progression, final learning outcome and educational degree classifications, even having significant impact on the employment and further professional career development. Furthermore, it is critical to maintain the academic QAA for the institutional reputation and has a positive impact on the learners' mentally well-being as well.
(5) There are 5 main recommendations proposed (**RQ5**) and discussed based on the reviewed outcome in Section 9, including technology readiness, e.g., theoretical computational modes, reliable datasets, quality of grading methods and evaluation of assessment accuracy, and so on. It may involve more than listed points within the context, because although research into ASAG is still at its early stage, it is a promising research area with impact on academic settings as well as commercially educational markets in the real-world.

Finally, Figure 23 shows a concept for our future expected automated grading system – **STPGE**.

S –Step 1 is to initiate the examination.
T – Step 2 is to take the examination, such as questions, sample answers set by the tutor, and student answers, which are put together in datasets.

P – In step 3, the student answers are processed by applying the techniques and post-processed text, and statistics are generated for evaluation.
G – In step 4, the processed texts are checked for their similarity level, and then, are graded.
E – In step 5, the differences between the human evaluation and the computed results are analysed to evaluate the accuracy and efficiency of ASAG at final stage.

It is clear that research into ASAG is still in its infancy, and only a small number of applications are available, most of them at a research level. Automated grading of short answers is a particularly promising research area and could be a strong social and economic impact on the huge educational markets in real-world.

References

[1] Research and Markets. 2018. Global artificial intelligence market in education sector. 2018–2022, Retrieved June 4, 2020 from researchandmarkets.com/reports/4522319/global-artificial-intelligence-market-in#pos-0

[2] Venkata V. Subrahmanyam and Swathi Kailasam. 2018. Artificial intelligence and its implications in education. In *Proceedings of the International Conference on Improved Access to Distance Higher Education Focus on Underserved Communities and Uncovered Regions*. IDEA–2018.

[3] Mohammed I. Younis and Maysam S. Hussein. 2015. Construction of an online examination system with resumption and randomization capabilities. *International Journal of Computing Academic Research* 4, 2 (2015), 62–82.

[4] Lu Joan. 2011. Mobile assessment system – MES. [Artefact]. Retrieved 17 November 2022 from eprints.hud.ac.uk/id/eprint/18621/

[5] Bonthu Sridevi, Sree S. Rama, and Krishna MHM Prasad. 2023. Improving the performance of automatic short answer grading using transfer learning and augmentation. *Engineering Applications of Artificial Intelligence* 123, Part A (2023), 1–8.

[6] Emiliano D. Gobbo, Alfonso Guarino, Barbara Cafarelli, and Luca Grilli. 2023. Automatic evaluation of open-ended questions for online learning. *A Systematic Mapping, Studies in Educational Evaluation* 77 (2023), 101258.

[7] Ellis B. Page. 1966. The imminence of grading essays by computer. *The Phi Delta Kappan* 47, 5 (1966), 238–243.

[8] Neslihan Süzen, Alexander Gorban, Jeremy Levesley, and Evgeny Mirkes. 2020. Automatic short answer grading and feedback using text mining methods. *Procedia Computer Science* 169 (2020), 726–743. DOI: 10.1016/j.procs.2020.02.171

[9] Joan Lu, Qiang Xu, Mike Joy, Simon McLenna, Gail Newton Gail, Artem Boyarchuk, James Robert, Yousef Muhammad, Dominic Williams, and Simon Fawcett. 2018a. A mobile learning technology used in teaching and learning in english primary schools. In *Proceedings of the International Conference on E-Learning*, e-Bus., EIS, and e-Gov. | EEE'18.

[10] Joan Lu, Qiang Xu, Mike Joy, Simon McLenna, Malgorzata Pankowska, Stuart Toddington, Artem Boyarchuk, and Guo Shulun. 2018. Wireless Response system for multidisciplinary teaching and learning – case studies. *Proceedings on the International Conference on Internet Computing (ICOMP'18)*. 108–113.

[11] Joan Lu, Qiang Xu, Mike Joy, Simon McLenna, Gail Newton Gail, Artem Boyarchuk, James Robert, Yousef Muhammad, Dominic Williams, and Simon Fawcett. 2018c. Use of a student response system in primary schools - an empirical study. *International Journal of e-Education, e-Business, e-Management and e-Learning* 9, 4 (2018), 324–330.

[12] Moodle, Question types - MoodleDocs. 2020. Retrieved October 9, 2020 from docs. moodle.org/39/en/Question_types

[13] Goutam Majumder, Partha Pakray, Alexander Gelbukh, and David Pinto. 2016. Semantic textual similarity methods, tools, and applications: A survey. *Computación Y Sistemas* 20, 4 (2016), 647–665. https://www.scielo.org.mx/pdf/cys/v20n4/1405-5546-cys-20-04-00647.pdf

[14] Emad F. Al-Shalabi. 2016. An automated system for essay scoring of online exams in arabic based on stemming techniques and levenshtein edit operations. *International Journal of Computer Science Issues* 13, 5 (2016), 45–50. DOI: 10.20943/01201605.4550

[15] Akeem N. Olowolayemo, Santhy David, and Teddy Mantoro. 2018. Short answer scoring in english grammar using text similarity measurement. *International Conference on Computing, Engineering, and Design*. (2018), 131–136.

[16] Yuwei Huang, Xi Yang, Fuzhen Zhuang, Lishan Zhang, and Shengquan Yu. 2018. Automatic chinese reading comprehension grading by LSTM with knowledge adaptation. *Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2018: Advances in Knowledge Discovery and Data Mining*. 118–129. DOI: 10.1007/978-3-319-93034-3_10

[17] Shih-hung Wu and Chun-yu Yeh. 2019. A short answer grading system in Chinese by CNN. In *Proceedings of the 2019 IEEE 10th International Conference on Awareness Science and Technology*. DOI: 10.1109/ICAwST.2019.8923528

[18] Leila Ouahrani and Djamel Bennouar. 2020. AR-ASAG an arabic dataset for automatic short answer grading 45 evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 2634–2643.

[19] Sahu Archana and Kumar P. Kumar Bhowmick. 2020. Feature engineering and ensemble-based approach for improving automatic short-answer grading performance. *IEEE Transactions on Learning Technologies* 13, 1 (2020), 77–90. DOI: 10.1109/TLT.2019.2897997

[20] Page J. Matthew, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2020. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372 (2020), 1–9. DOI: 10.1136/bmj.n71

[21] Jaclyn Broadbent, Stefanie Sharman, Ernesto Panadero, and Matthew Fuller-Tyszkiewicz. 2021. How does self-regulated learning influence formative assessment and summative grade? Comparing online and blended learners. *The Internet and Higher Education* 50 (2021), 100805.

[22] Joan Lu, Aswin Sundaram, Zhaozong Meng, Priya A, Gehao Lu andand, and John B. Stav. 2012. Mobile exam system – MES: Architecture for database management system, (chapter), learning with mobile technologies. *Handheld Devices, and Smart Phones: Innovative Methods* 272, Chapter 1 (2012), 1–20. DOI: 10.4018/978-1-4666-0936-5.ch001

[23] Joan Lu, Zhaozong Meng, Gehao Lu, and John B. Stav. 2010. A new approach in improving operational efficiency of wireless response system. In *Proceedings of the 10th IEEE International Conference on Computer and Information Technology*. 2676–2693. eprints.hud.ac.uk/10656/, accessed 17 November 2022.

[24] Katrina Perry, Kane Meissel, and Mary F. Hill. 2022. Rebooting assessment. *Exploring the Challenges and Benefits of Shifting from Pen-and-Paper to Computer in Summative Assessment.* Educational Research Review 12 March 2022.

[25] Joan Lu. 2011. Student response system (SRS)/wireless response system (WRS) – a next-generation student response system for academia and industry. [Artefact], Retrieved November 17, 2022 from eprints. hud.ac.uk/id/eprint/18619/

[26] Fawcett Simon. 2016. Internal report, St. Joseph Catholic School. Huddersfield, available in Retrieved 11 November 2022 from https://xdir.hud.ac.uk/2016/The%20use%20of%20a%20Wireless%20Response%20System

[27] Brightspace.com, accessed 20, December 2023.

[28] Joan Lu, Qiang Xu, Mike Joy, Simon McLenna, Gail Newton Gail, Artem Boyarchuk, James Robert, Yousef Muhammad, Dominic Williams, and Simon Fawcett. 2018. A mobile learning technology Used in teaching and learning in English Primary Schools. In *Proceedings of the International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government.* | EEE'18, Las Vegas, USA.

[29] Google, Google.co.uk/forms/about, accessed December 2021.

[30] Aslıhan Torkul, Aslıhan TÜFEKCİ, and Utku Köse. 2004. Web tabanli sinav sistemleri (web based examination systems). In *Proceedings of the the 1st International Conference on Informatics, Cesme, Turkey.*

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* 4171–4186.

[32] Malgorzata Pikies and Junade Ali. 2021. Analysis and safety engineering of fuzzy string matching algorithms. *ISA Transactions* 113 (2021), 1–8.

[33] Arya Prabhudesai and Ta. N. B. Duong. 2019. Automatic short answer grading using siamese bidirectional LSTM based regression. In *Proceedings of the 2019 IEEE International Conference on Engineering, Technology and Education.* DOI: 10.1109/TALE48000.2019.9226026

[34] Tiaoqiao Liu, Wenbiao Ding, Zhiwei Wang, Jiliang Tang, Gale Yan Huang, and Zitao Liu. 2019. Automatic short answer grading via multiway attention networks. *Artificial Intelligence in Education* (2019), 169–173. DOI: 10.1007/978-3-03023207-8_32

[35] Mohammed Qorich and Rajae El Ouazzani. 2025. Detection of artificial intelligence-generated essays for academic assessment integrity using large language models. *Expert Systems With Applications* 291, 128405 (2025), 1–16.

[36] Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R. Pisani, and Kathryn Turner. 2023. Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review. *Computers in Biology and Medicine* 155 (2023), 106649.

[37] Alec Cook and Oktay Karakuş. 2024. LLM-commentator: Novel fine-tuning strategies of large language models for automatic commentary generation using football event data. *Knowledge-Based Systems* 300 (2024), 112219.

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

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems.* 5998–6008.

[40] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 328–339.
[41] Sridevi Bonthu, Rama. S. Sree, and Prasad MHM Krishna. 2024. Framework for automation of short answer grading based on domain-specific pre-training. *Engineering Applications of Artificial Intelligence* 137, Part A (2024), 109163.
[42] Roshni M. Balakrishnan, Peeta B. Pati, Rimjhim P. Singh, S. Santhanalakshmi, and Priyanka Kumar. 2024. Fine-Tuned T5 for auto-grading of quadratic equation problems. *Procedia Computer Science* 235 (2024), 2178–2186.
[43] Qi Zhu, Yuxian Gu, Lingxiao Luo, Bing Li, Cheng Li, Wei Peng, Xiaoyan Zhu, and Minlie Huang. 2021. When does Further Pre-training MLM Help? An Empirical Study on Task-Oriented Dialog Pre-training, Proceedings of the Second Workshop on Insights from Negative Results in NLP, November 10, 2021. ©2021 Association for Computational Linguistics, ISBN 978-1-954085-93-0. 54–61.
[44] Xi Yang, Zhang Lishan, and Shengquan Yu. 2017. Can short answers to open response questions be auto graded without a grading rubric? *Proceedings of the AIED*. DOI: 10.1007/978-3-319-61425-0_72
[45] Shiyan Yang. 2020. Deep automated text scoring model based on memory network. In *Proceedings of the 2020 International Conference on Computer Vision, Image and Deep Learning*. 480–484. DOI: 10.1109/CVIDL51233.2020.00-46
[46] Gerd Kortemeyer. 2023. Performance of the pre-trained large language model GPT-4 on automated short answer grading. arXiv:2309.09338v1. Retrieved from https://arxiv.org/abs/2309.09338v1
[47] Da Wu, Jingye Yang, and Kai Wang. 2024. Exploring the reversal curse and other deductive logical reasoning in BERT and GPT-based large language models. *Patterns* 5, 9 (2024), 101030.
[48] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. arXiv:2302.13971. Retrieved from https://arxiv.org/abs/2302.13971
[49] Haoze Du, Qinjin Jia, Edward Gehringer, and Xianfang Wang. 2024. Harnessing large language models to auto-evaluate the student project reports. *Computers and Education: Artificial Intelligence* 7 (2024), 100268.
[50] D. M. Anisuzzaman, Jeffrey G. Malins, and Zachi I. Attia. 2025. Fine-tuning large language models for specialized use cases. *Mayo Clinic Proceedings: Digital Health*.
[51] Bernadette Quah, Yong Chee Weng, C. W. M. Lai, and I. Islam. 2024. Performance of large language models in oral and maxillofacial surgery examinations. *International Journal of Oral and Maxillofacial Surgery* 53, 10 (2024), 881–886. © 2024 International Association of Oral and Maxillofacial Surgeons. Published by Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.
[52] Hyeon Jo. 2023. Understanding AI tool engagement: A study of ChatGPT usage and word-of-mouth among university students and office workers. *Telematics and Informatics* 85 (2023), 102067.
[53] Qiang Xu and Joan Lu. 2023. Embracing ChatGPT in the teaching and learning of finite element analysis in engineering courses, 2023. In *Proceedings of the International Conference on Computational Science and Computational Intelligence*. | 979-8-3503-6151-3/23/∃31.00 ©2023 IEEE | DOI: 10.1109/CSCI62032.2023.00296
[54] Marcin Jukiewicz. 2024. The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process. *Thinking Skills and Creativitiy* 52 (2024), 101522.
[55] Yolanda Freire, Andrea S. Laorden, Jaime O. Pérez, Margarita G. Sánchez, Víctor D. García, and Ana Suárez. 2024. ChatGPT performance in prosthodontics: Assessment of accuracy and repeatability in answer generation. *The Journal of Prosthetic Dentistry* 131, 4 (2024), 659.e1–659.e6.
[56] Michael Haman and Milan Školník. 2024. Using ChatGPT to conduct a literature review. *Accountability and Research* 31, 8 (2024) 8.
[57] Yasir AlShehri, Mark McConkey, and Parth Lodhia. 2024. ChatGPT provides satisfactory but occasionally inaccurate answers to common patient hip arthroscopy questions. *Arthroscopy* 41, 5 (2025), 1337–1347. DOI: 10.1016/j.arthro.2024.06.017
[58] Luiz Rodrigues, Filipe D. Pereira, Luciano Temporal, Borges Cabral, Dragan Gasevic, G. L. Ramalho, and Rafael Mello. 2024. Assessing the quality of automatic-generated short answers using GPT-4. *Computers and Education: Artificial Intelligence* 7 (2024), 100248.
[59] Matthew L. Magruder, Ariel N. Rodriguez, Jason C. J. Wong, Orry Erez, Nicolas S. Piuzzi, Gil R. Scuderi, James D. Slover, Jason H. Oh, Ran Schwarzkopf, Antonia F. Chen, Richard Iorio, Stuart B. Goodman, and Michael A. Mont. 2024. Assessing ability for ChatGPT to answer total knee arthroplasty-related questions. *The Journal of Arthroplasty* 39, 8 (2024), 2022–2027.
[60] Aqdas Malik, M. Laeeq Khan, Khalid Hussain, Junaid Qadir, and Ali Tarhini. 2024. AI in higher education: unveiling academicians' perspectives on teaching, research, and ethics in the age of ChatGPT. *Interactive Learning Environments* 33, 3 (2025), 2390–2406.
[61] Varma Sashank, Emily M. Sanford, Vijay Marupudi, Olivia Shaffer, and Brook R. Lea. 2024. Recruitment of magnitude representations to understand graded words. *Cognitive Psychology* 153 (2024), 101673.

[62] Garg Jai, Jatin Papreja, Kumar Apurva, and Goonjan Jain. 2022. Domain-specific hybrid BERT based system for automatic short answer grading. In *Proceedings of the 2nd International Conference on Intelligent Technologies. Hubli, India*, 1–6. DOI: 10.1109/CONIT55038.2022.9847754

[63] Ghavidel Hadi, Amal Zouaq, and Michel Desmarais. 2020. Using BERT and XLNET for the automatic short answer grading task. *CSEDU 2020 - Proceedings of the 12th International Conference on Computer Supported Education* 1 (2020), 58–67.

[64] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: generalized autoregressive pretraining for language understanding. *Proceedings of the 33rd International Conference on Neural Information Processing Systems* Article No.: 517 (2019), 5753–57. https://proceedings.neurips.cc/paper_files/paper/2019

[65] Aurpaa, Tanjim Taharat, Kazi Noshin Farihab, Kawser Hossainb, Samiha Maisha Jebab, Md Shoaib Ahmedc, d, Md. Rawnak Saif Adibb, Farhana Islama, and Farzana Akter. 2024. Deep transformer-based architecture for the recognition of mathematical equations from real-world math problems. *Heliyon* 10, 20 (2024), e39089.

[66] Xinfeng Ye and Sathiamoorthy Manoharan. 2018. Machine learning techniques to automate scoring of constructed-response type assessments. In *Proceedings of the 2018 28th EAEEIE Annual Conference*. DOI: 10.1109/EAEEIE.2018.8534209

[67] Omar Nael, Youssef ELmanyalawy, and Nada Sharaf. 2022. AraScore: A deep learning-based system for Arabic short answer scoring. *Array* 13 (2022), 100109.

[68] Leila Ouahrani and Djamel Bennouar. 2019. A vector space-based approach for short answer grading system. In *Proceedings of the 19th International Arab Conference on Information Technology*. DOI: 10.1109/ACIT.2018.8672717

[69] Jes Kadupitiya, Surangika Ranathunga, and Gihan Dias. 2017. Sinhala short sentence similarity measures using corpus-based similarity for short answer grading. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing*. 44–53. Retrieved from http://aclanthology.lst.uni-saarland.de/W16-3705.pdf

[70] Farouk Mamdouh. 2019. Measuring sentences similarity: A survey. *Indian Journal of Science and Technology* 12, 25 (2019), 1–11. DOI: 10.17485/ijst/2019/v12i25/143977

[71] Anak Putri, Agung Ratna, Lea Santiar, Ihsan Ibrahim, Prima D. Purnamasari, Dyah L. Luhurkinanti, and Adisa Larasati. 2019. Latent semantic analysis and winnowing algorithm based automatic Japanese short essay answer grading system comparative performance. In *Proceedings of the 2019 IEEE 10th International Conference on Awareness Science and Technology*. DOI: 10.1109/ICAwST.2019.8923226

[72] Cagtay N. Tulu, Ozge Ozkaya, and Umut Orhan. 2021. Automatic short answer grading with sem space sense vectors and MaLSTM. *IEEE Access* 9 (2021), 19270–19280. DOI: 10.1109/ACCESS.2021.3054346

[73] Ahmed Magooda, Mohamed A. Zahran, Mohsen Rashwan, Hazem Raafat, and Magda B. Fayek. 2016. Vector based techniques for short answer grading. *International Florida Artificial Intelligence Research Society Conference* (2016), 238–243.

[74] Lucas B. Galhardi and Jacques Brancher. 2018. Machine learning approach for automatic short answer grading: A systematic review. In *Proceedings of the Advances in Artificial Intelligence. IBERAMIA 2018*, G. Simari, E. Fermé, F. Gutiérrez Segura, and J. Rodríguez Melquiades (Eds.). Lecture Notes in Computer Science, Vol 11238, Springer, Cham.

[75] Xiaoyan Zhang, Lufeng Cao, and Yipeng Yin. 2016. Individualized learning through MOOC: Online automatic test system based on genetic algorithm. *Proceedings of the 2016 International Conference on Intelligent Information Processing*. 1–6. DOI: 10.1145/3028842.3028855

[76] S. Marvaniya, Swarnadeep Saha, Tejas I. Dhamecha, P. Foltz, Renuka Sindhgatta, and Bikram Sengupta. 2018. Creating scoring rubric from representative student answers for improved short answer grading. In *Proceedings of the International Conference on Information and Knowledge Management*. 993–1002. DOI: 10.1145/3269206.3271755

[77] Sreevidhya V. Vadakkadath and Jayasree Narayanan. 2021. Short descriptive answer evaluation using word-embedding techniques. In *Proceedings of the 12th International Conference on Computing Communication and Networking Technologies*. 1–4. DOI: 10.1109/ICCCNT51525.2021.9579636

[78] Kumar Sachin, Soumen Chakrabarti, and Roy Shourya. 2017. Earth mover's distance pooling over siamese LSTMs for automatic short answer grading. *IJCAI International Joint Conference on Artificial Intelligence* 0 (2017), 20462052. DOI: 10.24963/ijcai.2017/284

[79] Diamel Guessoum, Moeiz Miraoui, and Chakib Tadj. 2015. Survey of semantic similarity measures in pervasive computing. *International Journal on Smart Sensing and Intelligent Systems* 8, 1 (2015), 125–158. DOI: 10.21307/ijssis-2017-752

[80] Nitesh Pradhan, Manasi Gyanchandani, and Rajesh Wadhvani. 2015. A review on text similarity technique used in IR and its application. *International Journal of Computer Applications* 120, 9 (2015), 29–34. DOI: 10.5120/212574109

[81] Robert W. Irving and Campbell Fraser. 1992. Two algorithms for the longest common subsequence ofthree (or more) strings. In *Proceedings of the 3rd Annual Symposium on Combinatorial Pattern Matching*. 214–229.

[82] Wael H. Gomaa and A. Fahmy Aly. 2013. A survey of text similarity approaches. *International Journal of Computer Applications* 68, 13 (2013), 13–18. DOI: 10.5120/11638-7118

[83] Kumar Yaman, Swati Aggarwal, Debanjan Mahata, Rajiv R. Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2019. Get IT scored using AutoSAS - an automated system for scoring short answers. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (2019), 9662–9669. DOI: 10.1609/aaai.v33i01.33019662

[84] Aqeel Aqeel. 2012. Textual similarity, kongens lyngby 2012 IMM-BSc-2012-16. *Technical University of Denmark*. Retrieved 6 October 2020, from http://www2.imm.dtu.dk/pubdb/edoc/imm6364.pdf

[85] Christopher N. Chung, BłaŻej Miasojedow, Michał Startek, and Anna Gambin. 2019. Jaccard/tanimoto similarity test and estimation methods for biological presence-absence data. In *Proceedings of the 14th International Symposium on Bioinformatics Research and Applications*. DOI: 10.1186/s12859-019-3118-5

[86] Sheetal A. Takale and Sushma S. Nandgaonkar. 2010. Measuring semantic similarity between words using web documents. *International Journal of Advanced Computer Science and Applications* 1, 4 (2010), 78–85. DOI: 10.14569/ijacsa.2010.010414

[87] Philip Resnik. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*.

[88] Thabet Slimani. 2013. Description and evaluation of semantic similarity measures approaches. *International Journal of Computer Applications* 80, 10 (2013), 25–33. DOI: 10.5120/13897-1851

[89] Yuhua Li, David McLean, Zuhair Bandar, James O'Shea, and Keeley Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering* 18, 8 (2006), 1138–1150.

[90] Ming C. Lee, Jia W. Chang, and Tung C. Hsieh. 2014. A grammar-based semantic similarity algorithm for natural language sentences. *The Scientific World Journal* 2014 (2014), 1–17. http://dx.doi.org/10.1155/2014/437162

[91] Gotama J. W. Putra and Takenobu Tokunaga. 2017. Evaluating text coherence based on semantic similarity graph. *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing. ACL Anthology*, DOI: 10.18653/v1/W17-2410

[92] Yuan Zhang, Chen Lin, and Min Chi. 2020. Going deeper: Automatic short-answer grading by combining student and question models. *User Modelling and User-Adapted Interaction* 30, 1 (2020), 51–80. DOI: 10.1007/s11257019-09251-6

[93] Weicheng Ma and Torsten Suel. 2016. Structural sentence similarity estimation for short texts. In *Proceedings of the 29th International Florida Artificial Intelligence Research Society Conference*. 232–37. nyuscholars.nyu.edu/en/publications/structural-sentence-similarity-estimation-for-short-texts. Accessed 17 November 2022.

[94] Farouk Mamdouh. 2020. Measuring text similarity based on structure and word embedding. *Cognitive Systems Research* 63 (2020), 1–10. DOI: 10.1016/j.cogsys.2020.04.002

[95] Sultan Md Arafat, Cristobal Salazar, and Tamara Sumner. 2016. Fast and easy short answer grading with high accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1070–1075. DOI: 10.18653/v1/n16-1123

[96] Masaki Uto and Uchida Yuto. 2020. Automated short-answer grading using deep neural networks and item response theory. In *Proceedings of the International Conference on Artificial Intelligence in Education*. 334–339. DOI: 10.1007/978-3-030-52240-7_61

[97] Stefano Menini, Sara Tonelli, Giovanni D. Gasperis, and Pierpaolo Vittorin. 2019. Automated short answer grading: A simple solution for a difficult task. In *Proceedings of the CEUR Workshop*.

[98] Azad Sushmita, Binglin Chen, Maxwell Fowler, Matthew West, and Craig Zilles. 2020. Strategies for deploying unreliable AI graders in high transparency high-stakes exams. In *Proceedings of the International Conference on Artificial Intelligence in Education*. 16–28. DOI: 10.1007/978-3-030-52237-7_2

[99] Yuan Zhang, Rajat Shah, and Min Chi. 2016. Deep learning + student modeling + clustering: A recipe for effective automatic short answer grading. In *Proceedings of the 9th International Conference on Educational Data Mining*. 562–567.

[100] Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. 2019. Improving short answer grading using transformer-based pre-training. In *Proceedings of the International Conference on Artificial Intelligence in Education*. DOI: 10.1007/978-3030-23204-7_39

[101] Xinhua Zhu, Wu Han, and Zhang Lanfang. 2022. Automatic short-answer grading via BERT-based deep neural 33 networks. In *IEEE Transactions on Learning Technologies* 15, 3 (2022), 364–375. DOI: 10.1109/TLT.2022.3175537

[102] Zaira H. Amur, Hooi Yew Kwang Soomro, and Gul Muhammad. 2022. Automatic short answer grading (ASAG) using attention-based deep learning MODEL. In *Proceedings of the International Conference on Digital Transformation and Intelligence*. 1–7. DOI: 10.1109/ICDI57181.2022.10007187

[103] Ifeanyi G. Ndukwe, Ben K. Daniel, and Chukwudi E. Amadi. 2019. A machine learning grading system using chatbots. In *Proceedings of the Artificial Intelligence in Education AIED 2019*, S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, and R. Luckin (Eds.). Lecture Notes in Computer Science, Vol 11626, Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-23207-8_67

[104] Lucija Petricioli, Skračić K. Petrović, Juraj and Pale Predrag. 2023. Exploring pre-scoring clustering for short answer grading. In *Proceedings of the 46th MIPRO ICT and Electronics Convention* 2014 (2023), 1567–1571. DOI: 10.23919/MIPRO57284.2023.10159981

[105] Rasha M. Badry, Mostafa Ali, Esraa Rslan, and Mostafa R. Kaseb. 2023. Automatic arabic grading system for short answer questions. In *IEEE Access* 11 (2023), 39457–39465. DOI: 10.1109/ACCESS.2023.3267407

[106] M. K. Vijaymeena and K. Kavitha. 2016. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal* 3, 1 (2016), 19–28. DOI: 10.5121/mlaij.2016.3103

[107] Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2010. Cross-language plagiarism detection. *Language Resources and Evaluation* 45, 1 (2010), 45–62. DOI: 10.1007/s10579-009-9114-z

[108] Franco-Salvador Marc, Parth Gupta, Paolo Rosso and Rafael E. Banchs. 2016. Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language. *Knowledge-Based Systems* 11, C (2016), 87–99. DOI: 10.1016/j.knosys.2016.08.004

[109] Jonathan Nau, Aluizio H. Filho, and Guilherme Passero. 2017. Evaluation semantic analysis methods for short answer grading using linear regression. *PEOPLE: International Journal of Social Science* 3, 2 (2017), 437–450.

[110] Nisha Varghese and M. Punithavalli. 2020. Semantic similarity analysis on knowledge based and prediction based models. *International Journal of Innovative Technology and Exploring Engineering* 9, 6 (2020).

[111] Fabian Zehner, Christine Sälzer, and Frank Goldhammer. 2016. Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement* 76, 2 (2016), 280–303. DOI: 10.1177/0013164415590022

[112] Bennouar Djame. 2017. An automatic grading system based on dynamic corpora. *International Arab Journal of Information Technology* 14, 4A (2017), 552–564. DOI: 10.1136/bmj.n160

[113] Hasanah Uswatun, Tri Astuti, Rizki Wahyudi, Zanuar Rifai, and Rilas A. Pambudi. 2018. An experimental study of text preprocessing techniques for automatic short answer grading in Indonesian. *Proceedings of the 2018 3rd International Conference on Information Technology, Information Systems and Electrical Engineering*. 230–234. DOI: 10.1109/ICITISEE.2018.8720957

[114] Hasanah Uswatun, Adhistya E. Permanasari, Sri S. Kusumawardani, and Feddy S. Pribadi. 2019. A scoring rubric for automatic short answer grading system. *Telkomnika (Telecommunication Computing Electronics and Control)* 17, 2 (2019), 763–770. DOI: 10.12928/TELKOMNIKA.V17I2.11785

[115] Saha Swarnadeep, Tejas I. Dhamecha, Marvaniya Smit, Sindhgatta Renuka, and Sengupta Bikram. 2018. Sentence level or token level features for automatic short answer grading?: Use both. In *Proceedings of the International Conference on Artificial Intelligence in Education*. 503–517. DOI: 10.1007/978-3-319-93843-1_37

[116] Ayse Çınar, Elif Ince, Murat Gezer, and Özgür Yılmaz. 2020. Machine learning algorithm for grading open-ended physics questions in turkish. *Education and Information Technologies* 25 (2020), 3821–3844. DOI: 10.1007/s10639-02010128-0

[117] Marvin C. Wijaya. 2021. Automatic short answer grading system in indonesian language using BERT machine learning. *Revue d'Intelligence Artificielle* 35, 6 (2021), 503–509. DOI: 10.18280/ria.350609

[118] Sarah Hassan, Aly A. Fahmy, and Mohammad El-Ramly. 2018. Automatic short answer scoring based on paragraph embeddings. *International Journal of Advanced Computer Science and Applications* 9, 10 (2018), 397–402. DOI: 10.14569/IJACSA.2018.091048

[119] Ifeanyi G. Ndukwe, Chukwudi E. Amadi, Larian M. Nkomo, and Ben K. Daniel. 2020. Automatic grading system using sentence-BERT network. In *Proceedings of the International Conference on Artificial Intelligence in Education*. 224–227. DOI: 10.1007/978-3-030-52240-7_41

[120] Chaturvedi Bhuvnesh and Rohini Basak. 2019. Automatic short-answer grading using corpus-based semantic 12 similarity measurements. In *Proceedings of the ICACIE-2019 International Conference*. DOI: 10.13140/RG.2.2.15125.88808

[121] Chaturvedi Bhuvnesh and Rohini Basak. 2021. Automatic short answer grading using corpus-based semantic similarity measurements. In *Proceedings of the Progress in Advanced Computing and Intelligent Engineering*, Springer Singapore. DOI: 10.1007/978-981-15-6353-9_24

[122] Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. 752–762. Retrieved from https://aclanthology.org/P11-1076.pdf

[123] Kaggle.com (2022). Retrieved November 17, 2022 from kaggle.com

[124] Chul Sung, Tejas Dhamecha, Swarnadeep Saha, Tengfei Ma, Vinay Reddy, and Rishi Arora. 2020. Pre-training BERT on domain resources for short answer grading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 6071–6075.

[125] Michael Mohler and Rada Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the ACL*. Association for Computational Linguistics, 567–575. Retrieved from https://aclanthology.org/E09-1065.pdf

[126] Andrew K. F. Lui, Sin C. Ng, and Stella W. N. Cheung. 2020. Entropy-based recognition of anomalous answers for efficient grading of short answers with an evolutionary clustering algorithm. In *Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence*. 3091–3098. DOI: 10.1109/SSCI47803.2020.9308137

[127] Beena Ahmed, Mohan Krishna, Kagita C. Wijenayake, and J. Ravishankar. 2019. Implementation guidelines for an automated grading tool to assess short answer questions on digital circuit design course. In *Proceedings of the 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering*. 1142–1145. DOI: 10.1109/TALE.2018.8615228

[128] Roy Shourya, Himanshu S. Bhatt, and Y. Narahari. 2016. Transfer learning for automatic short answer grading. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*. 1622–1623. DOI: 10.3233/978-1-61499-672-9-1622

[129] Roy Shourya, Y. Narahari, and Om Deshmukh. 2015. A perspective on computer assisted assessment techniques for short free-text answers. In *Proceedings of the International Computer Assisted Assessment Conference*, E. Ras and D. Joosten-ten Brinke (Eds.). Vol 571, Springer, Cham.

[130] Mohammed Azam Sayeed and Gupta Deepa. 2022. Automate descriptive answer grading using reference based models. *OITS International Conference on Information Technology* (2022), 262–267. DOI: 10.1109/OCIT56763.2022.00057

[131] David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. Educational testing service, A framework for evaluation and use of automated scoring. *Wiley, Educational Measurement: Issues and Practice* 31, 1 (2012), 2–13. DOI: https://doi.org/10.1111/j.1745-3992.2011.00223.x

[132] Abbirah Ahmed, Arash Joorabchi, and Martin J. Hayes. 2022. On the application of sentence transformers to automatic short answer grading in blended assessment. In *Proceedings of the 33rd Irish Signals and Systems Conference*. 1–6. DOI: 10.1109/ISSC55427.2022.9826194

[133] Yishan Chen, Jianhua Luo, Xinhua Zhu, Han Wu, and Shangbo Yuan. 2023. A cross-lingual hybrid neural network with interaction enhancement for grading short-answer texts. In *IEEE Access* 11 (2023), 37508–37514. DOI: 10.1109/ACCESS.2023.3260840

[134] BBC News. 2022. bbc.co.uk/news/newsbeat-58157807. Accessed 17 November 2022.

[135] Zeng Ling – Li, Huaning Wang, Panpan Hu, Bo Yang, Weidan Pu, Hui Shen, Xingui Chen, Zhening Liu, Hong Yin, Qingrong Tan, Kai Wang and Dewen Hu. 2018. Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. *EBioMedicine* 30 (2018), 74–85.

[136] Latif Ehsan and Xiaoming Zhai. 2024. Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence* 6 (2024) 100210.

[137] Mariano Maisonnave, Fernando Delbianco, Fernando Tohmé, Ana Maguitman, and Evangelos Milios. 2022. Detecting ongoing events using contextual word and sentence embeddings. *Expert Systems with Applications* 209 (2022), 118257.

[138] Zahid Halim and Zouq Aqsa. 2021. On identification of big-five personality traits through choice of images in a real-world setting. *Multimed Tools Appl* 80 (2021), 33377–33408. DOI: https://doi.org/10.1007/s11042-021-11419-5

[139] Zhaozong Meng and Joan Lu. 2011. Implementing the emerging mobile technologies in facilitating mobile exam system. In *Proceedings of the 2nd International Conference on Networking and Information Technology*. IPCSIT 25th-26th November 2011, IACSIT Press, Hong Kong, China, 80–88.

[140] Lishan Zhang, Yuwei Huang, Xi Yang, Shengquan Yu, and Fuzhen Zhuang. 2019. An automatic short-answer grading model for semi-open-ended questions. *Interactive Learning Environments* 30, 1 (2019), 177–190. DOI: 10.1080/10494820.2019.1648300

[141] Mieskes Margot and Padó Ulrike. 2018. Work smart–reducing effort in short-answer grading. *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning at SLTC 2018*. 57–68. eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) ISBN 978-91-7685-173-9.

[142] Juei-Yian Lin, Jhih-Yuan Huang, and Wei-Po Lee. 2021. Question-answer generation for data augmentation. In *Proceedings of the 2021 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*.

[143] Nicholas K. Corrêa, Sophia Falk, and Nythamar D. Oliveira. 2024. TeenyTinyLlama: Open-source tiny language models trained in Brazilian portuguese. *Machine Learning with Applications* 16, 100558 (2024), 1–12.

[144] Stanislav Chumakov, Anton Kovantsev, and Anatoliy Surikov. 2024. Ensuring accuracy and equity in vaccination information from ChatGPT and CDC: Mixed-methods cross-language evaluation. *JMIR Formative Research* 8 (2024), 11.

[145] Stanislav Chumakov, Anton Kovantsev, and Anatoliy Surikov. 2023. Generative approach to aspect based sentiment analysis with GPT language models. *Procedia Computer Science* 229 (2023), 284–293.

[146] Gerd Kortemeyer. 2024. Performance of the pre-trained large language model GPT-4 on automated short answer grading. *Discov Artif Intell* 4, 47 (2024), 1–6. DOI: https://doi.org/10.1007/s44163-024-00147-y

[147] He Zheng, Qing Sun, Qiushou Li, Yunxin Liu, Yuanxin Ouyang, and Qinghua Cao. 2025. FusionASAG: An LLM-enhanced automatic short answer grading model for subjective questions in online education. In *Proceedings of the Computer Science and Educational Informatization*. K. Zhang, X. Song, M. S. Obaidat, A. Bilal, J. Hu, Z. Lu, (eds), CSEI 2024 2024. Communications in Computer and Information Science, vol 2447. Springer, Singapore. DOI: https://doi.org/10.1007/978-981-96-3735-5_4

[148] Maximilian Tornqvist, Mosleh Mahamud, Erick Mendez Guzman, and Alexandra Farazouli. 2023. ExASAG: Explainable framework for automatic short answer grading. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 361–371.

[149] Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. 2023. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence* 6 (2024). 100213. DOI: 10.1016/j.caeai.2024.100213

[150] Zhengyang Xiao, Eunseo Lee, Sophia Yuan, Roland Ding, and Yinjie J. Tang. 2025. Generative AI in graduate bioprocess engineering exams: Is attention all students need? *Education for Chemical Engineers* 52 (2025), 133–140.

[151] Hongchen Wang, Kangming Li, Scott Ramsay, Yao Fehlis, Edward Kim, and Jason Hattrick-Simpers. 2025. Evaluating the performance and robustness of LLMs in materials science Q&A and property predictions. *Digital Discovery* 4, 6 (2025), 1612–1624.

Received 4 January 2023; revised 12 June 2025; accepted 19 June 2025