# Automatic Short Answer Grading (ASAG) In Education: A Review Of AI And Deep Learning Techniques

**Ashvi Patel[1*], Rupal Chaudhari[2], Mehul Patel[3]**

Masters In Computer Engineering Department / Sankalchand Patel University, Visnagar, India [1*]
Assistant. Professor, Computer Engineering Department / Sankalchand Patel University, Visnagar, India [2]
Assistant. Professor, Computer Engineering Department / Sankalchand Patel University, Visnagar, India [3]

ampbscit_spce@spu.ac.in, rrchaudharice_spce@spu.ac.in , mspatelit_spce@spu.ac.in

**Abstract:** Automatic Short Answer Grading (ASAG) has become a vital research area in educational technology, aiming to provide scalable, efficient, and fair evaluation of student responses. Early studies using machine learning demonstrated the feasibility of reducing grading workload through handcrafted features and statistical classifiers. However, such approaches struggled with semantic variability and domain adaptation. The introduction of deep learning enabled richer semantic representation, improving grading accuracy and robustness across tasks. In recent years, transformer based models have become the dominant paradigm. BERT and its variants, including Sentence-BERT and hybrid extensions, have consistently outperformed traditional neural networks by capturing deep contextual embeddings and semantic similarity more effectively. Comparative studies further confirm BERT's superiority over earlier embedding based and RNN based approaches, while also revealing challenges related to interpretability and domain transfer. More recent explorations into large language models, such as GPT and T5, demonstrate strong zero-shot and few-shot capabilities, extending the potential of ASAG but raising concerns around transparency, fairness, and multilingual support.This review synthesizes findings across two decades of research, emphasizing the evolution from feature-driven methods to BERT centered deep learning approaches and recent advances with LLMs. Open challenges remain in dataset scarcity, interpretability, multilingual grading, and trustworthy deployment in real world classrooms. The paper concludes by outlining future directions that integrate hybrid deep learning LLM approaches, benchmark development, and ethical frameworks to advance reliable and equitable ASAG.

**Keywords:** Automatic Short Answer Grading (ASAG) . Machine Learning . Recurrent Neural Networks (RNNs) . Convolutional Neural Networks (CNNs) . Deep Learning . Transformer Models . Bidirectional Encoder Representations from Transformers(BERT) . Generative Pre-trained Transformer (GPT) . Text-to-Text Transfer Transformer (T5) . Large Language Models (LLMs) . Educational Assessment.

## I. INTRODUCTION

The automation of assessment has emerged as a crucial research domain in educational technology, particularly with the exponential growth of digital learning environments. Among various types of assessments, Automatic Short Answer Grading (ASAG) stands out due to its potential to reduce teachers' workload while ensuring fairness, scalability, and timely feedback. Traditional manual grading is not only labor intensive but also subjective, often influenced by fatigue or bias. Consequently, researchers have sought to design intelligent systems that can evaluate student responses with high accuracy and consistency. Early efforts in ASAG were predominantly based on machine learning techniques with handcrafted features. For example, classification algorithms leveraging lexical, syntactic, and semantic features demonstrated that even simple models could partially reduce grading workloads [1]. However, these approaches suffered from limited generalization capabilities, as manually engineered features often failed to capture the deeper semantic nuances of natural language.

The rise of deep learning fundamentally transformed ASAG research. Pre-trained embeddings such as Word2Vec and GloVe initially provided better semantic representations, but their static nature restricted their ability to handle context dependent meanings [2]. The introduction of transformer based architectures marked a paradigm shift. In particular, BERT enabled models to capture bidirectional contextual dependencies, offering richer semantic representations than earlier embedding techniques [3,4]. These advancements led to significant performance improvements in automated grading tasks, with studies consistently reporting superior accuracy compared to traditional feature based models [5,7]. Recent research has increasingly focused on domain adaptation, hybrid architectures, and multilingual capabilities. For instance, BERT based frameworks have been extended into domain-specific variants to improve grading in specialized contexts [6], while hybrid approaches combining transformers with handcrafted features have shown further performance gains [13]. Comparative evaluations between transformer models, such as BERT, GPT, and T5, highlight that while GPT and T5 excel in generative tasks, BERT remains particularly effective for discriminative tasks like ASAG due to its robust contextual embeddings [15]. Moreover, multilingual and multi type answer grading systems have been explored to enhance the adaptability of ASAG systems in global education settings [10].

Another critical research direction involves the trustworthiness, interpretability, and workload reduction in automated grading. Studies have emphasized the importance of building systems that not only achieve high accuracy but also gain teachers' trust by providing transparent grading rationales [8,10,11]. Frameworks such as GradeAid have demonstrated the feasibility of implementing ASAG in real world classrooms, balancing automation with reliability [9]. In addition, recent work has begun to investigate the capabilities of LLMs such as GPT-4, showing promising yet mixed results in comparison to BERT based approaches [14]. Despite these advancements, challenges remain in terms of handling open domain responses, avoiding bias, scaling to diverse curricula, and ensuring interpretability. Therefore, reviewing the progression from traditional feature based models to deep learning and transformer based methods especially with a focus on BERT is critical for identifying research gaps and setting directions for future development.

In this context, a comprehensive review of ASAG research is both timely and necessary. By tracing the field's progression from feature engineered ML approaches to deep learning, and ultimately to transformer based architectures with BERT at the forefront, this paper provides a structured synthesis of how ASAG has evolved, the comparative strengths and weaknesses of different approaches, and the emerging directions that are likely to shape the future of automated assessment. This review thus aims to inform both researchers and practitioners by highlighting not only what has been achieved but also what critical challenges remain unresolved.

## 2. Background And Fundamentals Of ASAG

Automatic Short Answer Grading (ASAG) refers to the automated evaluation of short, open ended responses that typically range from one to three sentences. Unlike multiple choice questions, which test recognition, short answers are designed to measure deeper conceptual understanding and the ability of students to articulate knowledge in their own words. The primary motivation behind ASAG is to deliver grading that is consistent, scalable, and less labor intensive, while also supporting timely feedback in digital learning environments [1].

### 2.1 Challenges in Short Answer Grading

ASAG poses unique difficulties compared to traditional automatic assessment methods. Variability in language expression means that two answers can differ lexically while being semantically equivalent, making simple keyword matching insufficient. Additional challenges include domain dependency where models trained on one subject may not generalize well to others and the subjectivity inherent in human grading, which complicates the definition of a "gold standard" [2].

### 2.2 Illustrative Example of Scoring

Grading in ASAG is not binary but instead spans a range of correctness depending on semantic accuracy, conceptual completeness, and contextual relevance. Table 1 illustrates the typical scoring spectrum (0-5 scale) for a representative question.

TABLE. I:
EXAMPLE OF ASAG SCORING FOR A SHORT-ANSWER QUESTION

| | |
|---|---|
| Question | What is Photosynthesis? |
| Reference Answer | Photosynthesis is the process by which green plants use sunlight to convert carbon dioxide and water into glucose and oxygen. |
| Student Answer 1<br>Possible Score<br>Interpretation | *It is the process where plants make food using sunlight.*<br>5 (Correct)<br>Complete and semantically accurate. |
| Student Answer 2<br>Possible Score<br>Interpretation | *Plants use sunlight to prepare food, but oxygen is not involved.*<br>3 (Partially Correct)<br>Key idea present but missing essential detail. |
| Student Answer 3<br>Possible Score<br>Interpretation | *Photosynthesis means animals breathe oxygen from plants.*<br>1 (Incorrect/Misconception)<br>Confuses key concepts. |
| Student Answer 4<br>Possible Score<br>Interpretation | *I don't know.*<br>0 (Incorrect)<br>Attempt/Incorrect |

This example highlights two core aspects of ASAG. First, answers with different lexical forms but equivalent meaning should be graded consistently. Second, partially correct or contextually incorrect responses must be distinguished from fully correct ones to ensure fairness.

## 2.3 Evolution of ASAG Approaches

Early research primarily relied on feature engineering, where lexical, syntactic, and semantic overlap measures were combined with classifiers such as support vector machines (SVMs) or logistic regression [3]. While these methods demonstrated the feasibility of automation, their reliance on handcrafted features limited scalability and cross domain robustness.The adoption of deep learning techniques, particularly RNNs and CNNs, provided richer semantic representations but still struggled with context dependent meanings when using static embeddings like Word2Vec and GloVe [4]. The emergence of transformer based models, especially BERT, marked a paradigm shift by enabling bidirectional contextual encoding. Numerous studies have since demonstrated BERT's superior performance over both feature based and earlier deep learning approaches in ASAG tasks [5].

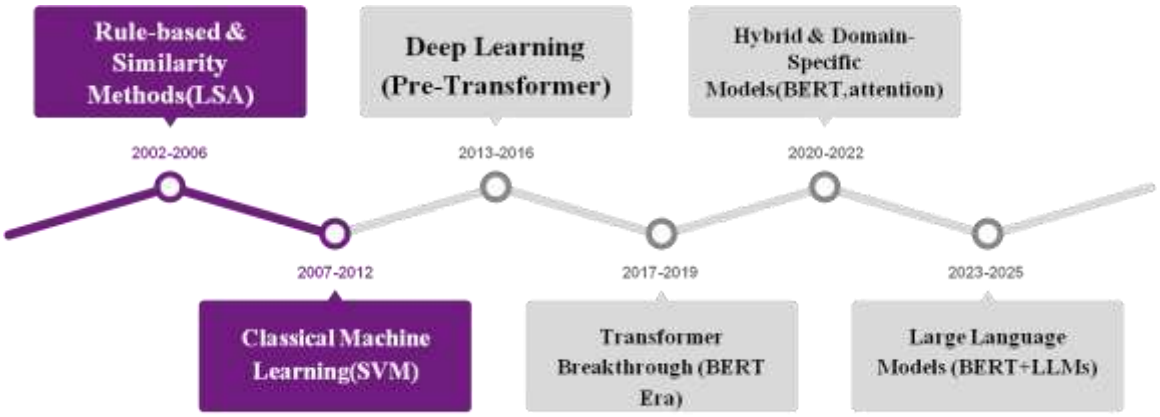


Fig. 1. Evolutionary Timeline of Automatic Short Answer Grading (ASAG).

## 2.4 General Pipeline of ASAG Systems

Although implementations vary, a generalized ASAG workflow can be described in sequential stages, as depicted in Figure 1. The process begins with the collection of student responses, which undergo preprocessing (tokenization, normalization, and sometimes domain-specific adaptations). The responses are then transformed into vector representations using embeddings or contextual encoders. These representations are processed by a predictive model (machine learning, deep learning, or transformers), which generates a numerical score. In advanced systems, the score may be accompanied by feedback or justification to enhance interpretability.

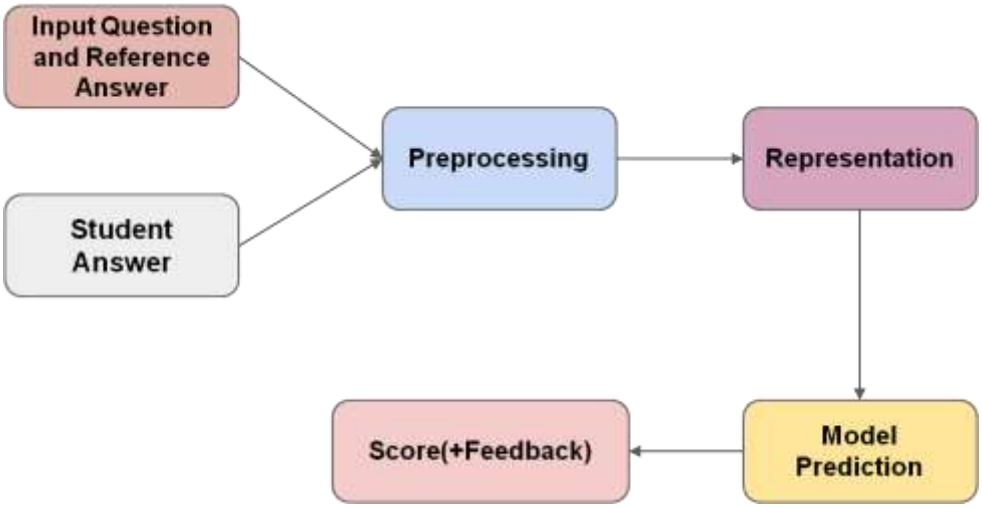


Fig. 2. Generalized pipeline of an Automatic Short Answer Grading system.

# III. LITERATURE REVIEW

ASAG has emerged as a significant area of educational technology research, aiming to enhance scalability and consistency in assessment. Its progression reflects a broader trend in Natural Language Processing and artificial intelligence: transitioning from rule based and machine learning techniques toward deep learning and transformer based models. This section reviews key contributions in the field, critically synthesizes their findings, and highlights gaps and challenges that inform the present study.

The foundation of ASAG research lies in traditional machine learning techniques, where systems depended heavily on handcrafted linguistic features such as lexical overlap, n-grams, or syntactic similarity [1]. While these methods demonstrated feasibility, their reliance on feature engineering limited adaptability across domains and datasets. For instance, models often failed when applied to unseen subject areas, raising concerns about generalizability. Nevertheless, this early phase was instrumental in demonstrating the potential of automation to reduce manual grading workload. Later comparative evaluations revealed that transfer learning and pretrained models offered substantial improvements over handcrafted features, particularly in capturing semantic meaning [2]. These studies suggested that machine learning was insufficient when applied in isolation, paving the way for the adoption of deep learning and contextual embeddings.

The integration of pretrained embeddings such as BERT and sentence-BERT represented a paradigm shift. Sentence-BERT demonstrated strong potential for semantic similarity tasks, including ASAG, by encoding responses into dense vector representations [3]. Similarly, comparative evaluations of transformer architectures confirmed the superiority of contextual embeddings in capturing meaning beyond surface level lexical matches [4]. These works highlighted that pretrained transformers reduced dependence on manual feature engineering and improved grading accuracy across diverse datasets. However, these advancements also surfaced critical challenges: pretrained models require substantial computational resources, are data hungry, and risk overfitting when applied to small, domain-specific datasets. These limitations highlight the necessity of balancing performance with practical implementation in educational environments.

Deep learning models broadened the methodological toolkit for ASAG, enabling sophisticated architectures to process student responses. Surveys and experimental studies identified strategies ranging from convolutional and recurrent neural networks to

hybrid systems [5]. While these models demonstrated improvements in semantic understanding, they often lacked transparency and raised concerns about explainability. Hybrid approaches offered a potential solution. By combining rule based or domain specific heuristics with transformer based embeddings, systems achieved a balance between interpretability and accuracy [6]. Such approaches not only enhanced performance in specialized tasks but also provided educators with partial insight into the grading rationale. Yet, hybrid systems risk becoming overly complex and difficult to generalize, an issue that continues to limit their adoption in large scale educational contexts.

The emergence of BERT marked a turning point for ASAG. BERT based grading models demonstrated significant gains in accuracy and reliability compared to earlier embedding and deep learning techniques [7]. Empirical evidence showed that contextualized word embeddings allow models to capture nuanced semantic differences in short answers, making them highly effective for automated assessment. Reviews consolidating prior research further emphasized that transformers dominate the field, outperforming traditional models in nearly every benchmark [8]. More recently, frameworks have been developed to operationalize ASAG in educational contexts. For example, modular systems demonstrated that transformers could be integrated into institutional platforms for practical deployment [9]. At the same time, studies on multilingual and diverse answer datasets revealed the challenges of trustworthiness and fairness, as models occasionally misgraded non standard but valid responses [10]. These findings underscore that while transformer based systems advance accuracy, they still grapple with inclusivity and robustness in real world settings.

TABLE. II:
COMPARISON OF EXISTING RESEARCH IN ASAG

| Year | Approach / Model | Key Contribution | Strengths | Limitations |
|---|---|---|---|---|
| 2018[1] | Traditional ML (lexical, n-grams, syntactic features) | Demonstrated feasibility of ASAG using handcrafted linguistic features | Simple, interpretable | Limited generalizability, heavy feature engineering |
| 2020[4] | Transformer models (BERT, sentence-BERT) | Established transformer superiority in semantic understanding | High accuracy, contextualized embeddings | Data hungry, resource intensive |
| 2021[6] | Hybrid transformer + rule-based | Combined BERT with heuristics for explainability | Balanced accuracy and interpretability | Overly complex, hard to generalize |
| 2022[10] | Multilingual ASAG with transformers | Evaluated fairness across diverse learner groups | Inclusivity, cross-lingual | Inconsistent grading, fairness challenges |
| 2023[12] | ASAG for comprehension (aphorism interpretation) | Applied deep learning to creative comprehension | Domain flexibility | Lacks interpretability |
| 2024[14] | GPT-4 for ASAG | Tested LLMs for grading tasks | Strong semantic adaptability | Bias, inconsistency, black-box |
| 2025[17] | Large Language Models (GPT-4, LLaMA, T5-XXL) for ASAG | valuated adaptability of LLMs for real-world classroom grading | High flexibility, multilingual grading, strong semantic reasoning | Bias, lack of transparency, high resource demand |

Recent years have seen a diversification of ASAG applications. Deep learning based systems have been applied to comprehension assessment tasks, such as evaluating student interpretations of aphorisms, demonstrating flexibility across different question types [12]. Hybrid approaches that combine deep learning with traditional features also continue to show promise in balancing interpretability with performance [13]. The most notable development has been the emergence of LLMs. Evaluations of GPT-4 for ASAG suggest strong potential, with the model demonstrating impressive semantic understanding and adaptability [14]. Yet, concerns remain about bias, inconsistency, and lack of transparency in grading. Comparative studies of BERT, GPT, and T5 further highlight that no single transformer dominates across all contexts. BERT often excels in fine-tuned tasks with limited data, while GPT demonstrates generative flexibility, and T5 offers task specific adaptability [15]. These comparisons suggest that model selection should be guided by the assessment context, rather than relying on a one-size-fits all approach.

ASAG research has also extended into related domains, confirming the transferability of approaches. For example, fine-tuned BERT models for classifying learner reviews demonstrated adaptability of these methods to broader educational NLP tasks [17]. This extension suggests that innovations in ASAG have the potential to influence other areas of educational technology, such as automated feedback generation, personalized learning systems, and large scale learner analytics.

Across the literature, several patterns emerge. First, there is a clear shift from feature based machine learning to pretrained transformers, with BERT establishing itself as a benchmark. However, while these models excel in accuracy, they often fall short in interpretability, fairness, and scalability. Second, hybrid approaches demonstrate promise in balancing performance with transparency, yet they remain limited in real world adoption due to complexity. Third, the rise of LLMs opens exciting possibilities but simultaneously raises concerns regarding bias, reliability, and lack of explainable reasoning.

## IV. CONCLUSION

The review of existing research on Automatic Short Answer Grading (ASAG) reveals a steady evolution from early machine learning techniques to advanced deep learning and transformer based approaches, with models such as BERT showing notable gains in semantic comprehension and grading performance. Despite these advancements, persistent challenges remain, particularly the scarcity of large, diverse, and multilingual datasets, as well as the lack of transparency, interpretability, and fairness in grading systems. Hybrid models and recent investigations into large language models like GPT offer promising directions, yet they introduce concerns regarding explainability, bias, and alignment with pedagogical standards. Moving forward, research should emphasize the development of standardized benchmark datasets, robust evaluation metrics, and interpretable frameworks, while also exploring adaptive, cross lingual, and domain generalizable ASAG systems. Integrating these innovations with intelligent tutoring systems and real-world classroom settings represents a crucial future direction for achieving both technological effectiveness and educational impact.

## REFERENCES

[1] Krithika, R., and Jayasree Narayanan. "Learning to grade short answers using machine learning techniques." Proceedings of the Third International Symposium on Women in Computing and Informatics. 2015.

[2] Gaddipati, Sasi Kiran, Deebul Nair, and Paul G. Plöger. "Comparative evaluation of pretrained transfer learning models on automatic short answer grading." arXiv preprint arXiv:2009.01303 (2020).

[3] Ndukwe, Ifeanyi G., et al. "Automatic grading system using sentence-BERT network." International Conference on Artificial Intelligence in Education. Cham: Springer, 2020.

[4] Camus, Leon, and Anna Filighera. "Investigating transformers for automatic short answer grading." International Conference on Artificial Intelligence in Education. Cham: Springer, 2020.

[5] Ahmed, Abbirah, Arash Joorabchi, and Martin J. Hayes. "On Deep Learning Approaches to Automated Assessment: Strategies for Short Answer Grading." CSEDU (2) (2022): 85-94.

[6] Garg, Jai, et al. "Domain-specific hybrid bert based system for automatic short answer grading." 2022 2nd International Conference on Intelligent Technologies (CONIT). IEEE, 2022.

[7] Zhu, Xinhua, Han Wu, and Lanfang Zhang. "Automatic short-answer grading via BERT-based deep neural networks." IEEE Transactions on Learning Technologies 15, no. 3 (2022): 364-375.

[8] Haller, Stefan, et al. "Survey on automated short answer grading with deep learning: from word embeddings to transformers." arXiv preprint arXiv:2204.03503 (2022).

[9] Del Gobbo, Emiliano, Alfonso Guarino, Barbara Cafarelli, and Luca Grilli. "GradeAid: a framework for automatic short answers grading in educational contexts—design, implementation and evaluation." Knowledge and Information Systems 65, no. 10 (2023): 4295-4334.

[10] Schneider, Johannes, Robin Richner, and Micha Riser. "Towards trustworthy autograding of short, multi-lingual, multi-type answers." International Journal of Artificial Intelligence in Education 33, no. 1 (2023): 88-118.

[11] Weegar, Rebecka, and Peter Idestam-Almquist. "Reducing workload in short answer grading using machine learning." International Journal of Artificial Intelligence in Education 34, no. 2 (2024): 247-273.

[12] Mardini, G., Ivan D., et al. "A deep-learning-based grading system (ASAG) for reading comprehension assessment by using aphorisms as open-answer-questions." Education and Information Technologies 29.4 (2024):

4565-4590.

[13] Kaya, Mustafa, and Ilyas Cicekli. "A hybrid approach for automated short answer grading." IEEE Access 12 (2024): 96332-96341.

[14] Kortemeyer, Gerd. "Performance of the pre-trained large language model GPT-4 on automated short answer grading." Discover Artificial Intelligence 4, no. 1 (2024): 47.

[15] Zaki, Muhammad Zayyanu. "Revolutionising translation technology: A comparative study of variant transformer models–BERT, GPT and T5." Computer Science and Engineering–An International Journal 14.3 (2024): 15-27.

[16] Chaudhari, Rupal, and Manish Patel. "Deep Learning in Automated Short Answer Grading: A Comprehensive Review." ITM Web of Conferences. Vol. 65. EDP Sciences, 2024.

[17] Chen, Xieling, et al. "Automatic Classification of Online Learner Reviews Via Fine-Tuned BERTs." International Review of Research in Open and Distributed Learning 26.1 (2025): 57-79.

[18] Jung, Ji Yoon, Lillian Tyack, and Matthias von Davier. "Combining machine translation and automated scoring in international large-scale assessments." Large-scale Assessments in Education 12, no. 1 (2024): 10.

[19] Jing, Shumin, O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, and A. Merceron. "Automatic Grading of Short Answers for MOOC via Semi-supervised Document Clustering." In EDM, pp. 554-555. 2015.

[20] Zhang, Yuan, Rajat Shah, and Min Chi. "Deep Learning+ Student Modeling+ Clustering: A Recipe for Effective Automatic Short Answer Grading." International Educational Data Mining Society (2016).

[21] Jiang, Lan, and Nigel Bosch. "Short answer scoring with GPT-4." In Proceedings of the Eleventh ACM Conference on Learning@ Scale, pp. 438-442. 2024.

[22] Chamieh, Imran, Torsten Zesch, and Klaus Giebermann. "Llms in short answer scoring: Limitations and promise of zero-shot and few-shot approaches." In Proceedings of the 19th workshop on innovative use of nlp for building educational applications (bea 2024), pp. 309-315. 2024.

[23] Tulu, Cagatay Neftali, Ozge Ozkaya, and Umut Orhan. "Automatic short answer grading with SemSpace sense vectors and MaLSTM." IEEE Access 9 (2021): 19270-19280.

[24] Balaha, Hossam Magdy, and Mahmoud M. Saafan. "Automatic exam correction framework (aecf) for the mcqs, essays, and equations matching." IEEE Access 9 (2021): 32368-32389.

[25] Sychev, Oleg, Anton Anikin, and Artem Prokudin. "Automatic grading and hinting in open-ended text questions." Cognitive Systems Research 59 (2020): 264-272.

[26] Bennouar, Djamal. "An automatic grading system based on dynamic corpora." Int. Arab J. Inf. Technol. 14, no. 4A (2017): 552-564.

[27] Liu, Tianyi, Julia Chatain, Laura Kobel-Keller, Gerd Kortemeyer, Thomas Willwacher, and Mrinmaya Sachan. "AI-assisted automated short answer grading of handwritten university level mathematics exams." arXiv preprint arXiv:2408.11728 (2024).

[28] Bernard, Jason, Ranil Sonnadara, Anthony N. Saraco, Josh P. Mitchell, Alex B. Bak, Ilana Bayer, and Bruce C. Wainman. "Automated grading of anatomical objective structured practical examinations using decision trees: An artificial intelligence approach." Anatomical Sciences Education 17, no. 5 (2024): 967-978.

[29] Dadu, Niharika, Harsh Vardhan Singh, and Romi Banerjee. "Grade Guard: A Smart System for Short Answer Automated Grading." arXiv preprint arXiv:2504.01253 (2025).

[30] Meyer, Gérôme, Philip Breuer, and Jonathan Fürst. "Asag2024: A combined benchmark for short answer grading." In Proceedings of the 2024 on ACM Virtual Global Computing Education Conference V. 2, pp. 322-323. 2024.