Int. J. of AI and Education Technology

♦ ISSN: 2835-2432 Vol: 05, No. 01, 2026 ♦ PP. 25–30

Article DOI: https://doi.org/10.54216/IJAIET.050104



REVIEW ARTICLE

A Systematic Review of AI-Powered Uzbek Short-Answer Grading Using NLP and Teacher-Annotated Datasets

Sanjar Raximjonov[1,*] **Eugene Q. Castro**[1]

[1] Department of Computer Science, Central Asian University, Tashkent, Uzbekistan

Emails: 220304@centralasian.uz · e.castro@centralasian.uz

Received: November 04, 2025 Revised: December 11, 2025 Accepted: January 18, 2026 ⋆ Corresponding author

ABSTRACT

This paper presents a Systematic Literature Review (SLR) of AI-powered automated short-answer grading, with a particular focus on low-resource languages such as Uzbek. The review follows the PRISMA 2020 guidelines to ensure transparency and methodological rigor. Relevant peer-reviewed studies published between 2018 and 2025 were systematically identified, screened, and analyzed across multiple academic databases. In total, 33 studies were included in the final synthesis. The reviewed literature indicates that transformer-based models, including mBERT and XLM-R, generally achieve stronger performance than traditional machine learning approaches, while recent large language models show potential in few-shot and zero-shot grading scenarios. The findings also highlight that the limited availability of teacher-annotated datasets remains a major challenge for developing reliable automated grading systems in low-resource educational contexts.

Keywords: Automated Short-Answer Grading ▪ Natural Language Processing ▪ Transformer Models ▪ Low-Resource Languages ▪ Uzbek Language ▪ Systematic Literature Review

1. INTRODUCTION

The rapid growth of digital education has increased the demand for scalable, consistent, and reliable assessment methods. In many educational systems, including Uzbekistan, short-answer questions are widely used to evaluate students' conceptual understanding and reasoning skills. However, grading such responses is still predominantly performed manually, which is time-consuming, subjective, and difficult to scale for large classes. These limitations often lead to inconsistencies in grading quality and delays in feedback delivery, which can negatively affect the learning process [1, 2].

Recent advances in artificial intelligence (AI) and natural language processing (NLP) have enabled the development of automated assessment systems capable of analyzing and scoring textual responses. In high-resource languages such as English, automated short-answer grading systems based on machine learning and deep learning techniques have demonstrated strong performance and high agreement with human graders. In particular, transformer-based models, including BERT and its multilingual variants, have significantly improved semantic understanding by capturing contextual representations of student responses [3].

Despite these advancements, low-resource languages such as Uzbek remain largely underexplored in the context of automated short-answer grading. One of the primary challenges is the limited availability of teacher-annotated datasets, which are essential for training and evaluating supervised learning models. Furthermore, the morphological richness of the Uzbek language and the lack of standardized evaluation benchmarks make it difficult to directly transfer models developed for high-resource languages.

Existing studies have proposed a wide range of automated grading approaches, including traditional machine learning

© 2026 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

methods, transformer-based architectures, and, more recently, large language models. However, differences in datasets, experimental setups, and evaluation metrics across studies make it challenging to draw consistent conclusions regarding model effectiveness, particularly for low-resource educational contexts.

To address these challenges, this study conducts a Systematic Literature Review (SLR) following the PRISMA 2020 guidelines. The objective of this review is to systematically identify, analyze, and synthesize existing research on AI-powered automated short-answer grading. Specifically, this review aims to examine supervised NLP models used for short-answer grading, analyze the characteristics of teacher-annotated datasets with an emphasis on low-resource languages, and identify key research gaps relevant to the development of an AI-based Uzbek short-answer grading assistant.

2. METHODOLOGY

This study follows a Systematic Literature Review (SLR) methodology in accordance with the PRISMA 2020 guidelines to ensure transparency, methodological rigor, and reproducibility [7]. The review protocol was defined in advance and applied consistently throughout the study selection and analysis process.

2.1 Research Design

The SLR approach was selected to provide a structured and unbiased synthesis of existing research on AI-powered automated short-answer grading. This methodology enables the systematic identification, evaluation, and comparison of prior studies while minimizing selection bias. The review was guided by the following research objectives: (i) to identify the dominant NLP-based models used for automated short-answer grading, (ii) to analyze the characteristics of teacher-annotated datasets, particularly in low-resource languages, and (iii) to identify key research gaps relevant to the Uzbek language context.

2.2 Research Questions

This Systematic Literature Review is guided by the following research questions:

- RQ1: What NLP-based models are most commonly used for automated short-answer grading?
- RQ2: What types of teacher-annotated datasets are used, particularly in low-resource languages?
- RQ3: What challenges and research gaps exist in applying automated short-answer grading to the Uzbek language context?

2.3 Data Sources and Search Strategy

A comprehensive and reproducible literature search was conducted across five recognized scholarly databases: IEEE Xplore, ACM Digital Library, ScienceDirect, JSTOR, and Google Scholar. The search was performed between **November 3 and November 5, 2025**, and targeted peer-reviewed studies published between 2018 and 2025 in order to capture recent advances in automated short-answer grading.

To ensure transparency and reproducibility, explicit Boolean search strings were defined and applied consistently across all databases, with minor syntactic adaptations made only where required by database-specific query rules (e.g., field tags or quotation handling).

The primary search string used in this review was: *("automatic short-answer grading" OR "automated assessment") AND ("natural language processing" OR "NLP" OR "transformer models") AND ("education" OR "student responses") AND ("low-resource languages" OR "multilingual").*

In databases that did not fully support complex Boolean expressions (e.g., Google Scholar), simplified variants of the same keywords were used while preserving the semantic intent of the original query. No additional conceptual terms were introduced. The same inclusion period, filters, and relevance criteria were applied across all databases to ensure comparability of the retrieved results.

2.4 Search Log

To ensure transparency and reproducibility, a detailed search log was maintained for each database. Table 2 summarizes the databases searched, exact search strings, applied filters, and the number of results retrieved.

2.5 Study Selection Process

The initial database search yielded 1,194 records. Duplicate entries were removed prior to screening. Titles and abstracts were then reviewed to assess relevance based on the inclusion and exclusion criteria. Full-text screening was subsequently performed to determine eligibility. As a result of this process, 33 studies were selected for the final review. The complete study selection procedure is summarized using a PRISMA flow diagram, as shown in Figure 1.

2.6 Quality Assessment

To ensure the reliability and relevance of the selected studies, a quality assessment was conducted during the full-text screening phase. Each study was evaluated based on the clarity of its methodology, relevance to automated short-answer grading, description of datasets, and reporting of evaluation metrics. Studies that did not meet these quality criteria were excluded from the final synthesis.

2.7 Data Extraction and Analysis

Relevant information was systematically extracted from each selected study using a structured data extraction framework. Extracted data included publication year, model architecture, dataset size, language, evaluation metrics, and reported performance. The extracted data were analyzed using both quantitative summarization and qualitative thematic synthesis to identify prevailing trends, methodological patterns, and limitations across the reviewed literature.

3. RESULTS

This section presents the synthesized results of the Systematic Literature Review based on the analysis of the 33 selected studies. The findings are organized into quantitative summaries and qualitative observations to highlight dominant research trends in automated short-answer grading.

Table 1. Inclusion and Exclusion Criteria

Criteria Type	Description
Inclusion Criteria	• Empirical studies on automated or AI/NLP-based short-answer grading. • Use of supervised learning with teacher- or human-annotated datasets. • Peer-reviewed journal or conference papers published between 2018 and 2025. • Studies reporting evaluation metrics (e.g., Accuracy, F1-score, Cohen's κ, MAE, RMSE). • Research conducted in educational contexts, including low-resource or morphologically rich languages.
Exclusion Criteria	• Review papers, theoretical works, or non-empirical studies. • Studies not focused on short-answer grading (e.g., essay-only or multiple-choice assessment). • Rule-based systems without AI or NLP modeling. • Duplicate or secondary publications based on the same dataset.

Table 2. Search Log per Database

Database	Date	Exact Search String	Filters Applied	Results
ScienceDirect	Nov 3, 2025	"automated short answer grading" OR "short answer scoring" AND (NLP OR transformer) AND ("teacher-annotated" OR "human-annotated")	2018–2025; English; Article/Conference	14
IEEE Xplore	Nov 3, 2025	("short answer grading" OR "automated assessment") AND (NLP OR BERT OR transformer)	2018–2025; English; Journals & Conferences	58
ACM Digital Library	Nov 4, 2025	("short answer" AND grading) AND (BERT OR transformer)	2018–2025; Full Text	553
JSTOR	Nov 4, 2025	"automated grading" AND education AND "short answer"	2018–2025; Education; Journals	4
ACM Digital Library	Nov 4, 2025	"NLP" AND "automated grading" AND education	2018–2025; All content	134
Mendeley	Nov 4, 2025	"teacher annotated dataset" AND "short answer"	2018–2025; Journal/Conference	231
Google Scholar	Nov 5, 2025	"AI short answer grading" "Uzbek" OR "low-resource" OR "teacher-annotated"	First 200 results screened	200

3.1 Study Summary

Table 3 summarizes the primary model categories used in the reviewed studies. Each study was categorized according to its dominant modeling approach, even when multiple models were evaluated.

Table 3. Summary of Included Studies by Primary Model Type

Category	Primary Models	Studies
Traditional ML	SVM, Logistic Regression	7
Transformer-based	BERT, mBERT, XLM-R	18
Large Language Models	GPT-based	8
Total		**33**

3.2 Datasets Used

The reviewed studies employed a variety of teacher-annotated datasets for automated short-answer grading. Most datasets were relatively small in scale, typically containing fewer than 1,000 annotated student responses. The majority of datasets were in English, while a limited number of studies focused on multilingual or low-resource language settings. Annotations were primarily provided by subject-matter experts or trained educators using predefined scoring rubrics. These datasets were commonly collected in educational contexts such as classroom assessments, exams, and short-answer evaluation tasks. No new datasets were collected or created in this study; all dataset information was extracted from the original peer-reviewed articles included in the review.

3.3 Quantitative Overview

The reviewed studies predominantly employ supervised learning approaches for automated short-answer grading. Transformer-based architectures have become the dominant modeling paradigm in recent research, while most datasets reported in the literature contain fewer than 1,000 teacher-annotated responses. This reflects the limited availability of labeled data, particularly in low-resource language settings. Commonly reported evaluation metrics include accuracy, Quadratic Weighted Kappa (QWK), and F1-score, which are widely adopted from both short-answer and automated essay scoring research [15, 12, 5].

3.4 Model Performance Trends

Across the reviewed literature, transformer-based models consistently outperform classical machine learning approaches such as logistic regression and support vector machines. Multilingual transformer models demonstrate improved performance in cross-lingual and low-resource scenarios by leveraging shared representations across languages. In addition, recent studies indicate that large language models can achieve competitive performance in few-shot and zero-shot grading settings. These results are reported findings from prior studies and are used for comparative analysis rather than direct experimental validation in this review [3, 4, 6].

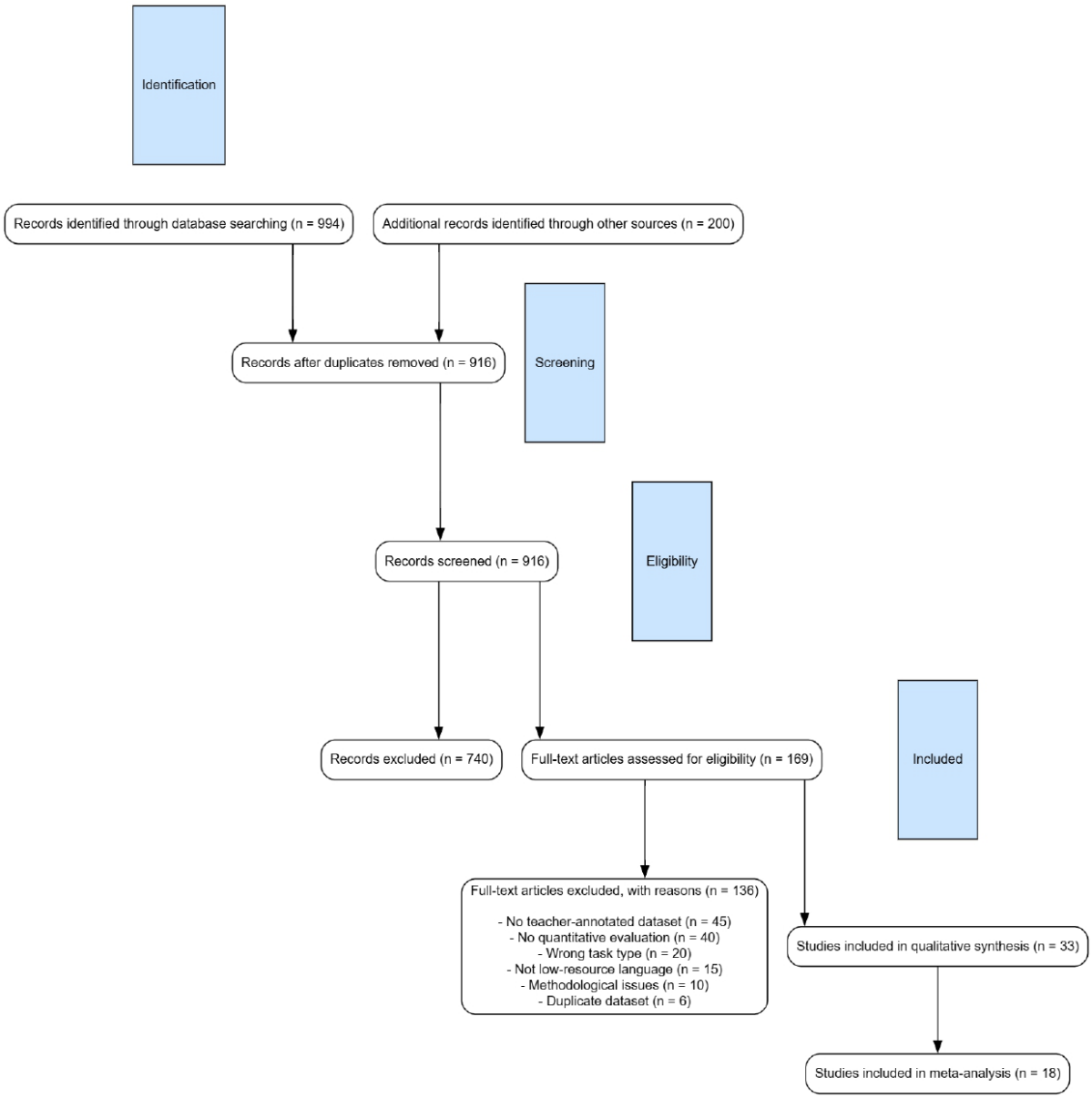


Figure 1. PRISMA 2020 flow diagram illustrating the study selection process.

3.5 Model Distribution

To provide an overview of methodological trends, Figure 2 illustrates the distribution of model categories employed in the reviewed studies. Transformer-based approaches consti-tute the largest proportion, followed by traditional machine learning methods and emerging large language model-based approaches. This distribution highlights the gradual shift in the research community toward more expressive neural architectures for automated short-answer grading.

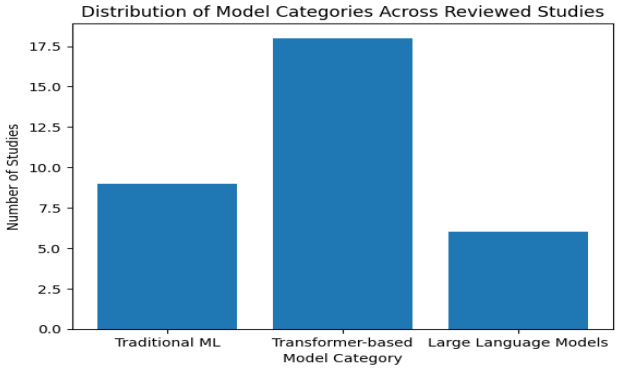


Figure 2. Distribution of model types across the reviewed studies.

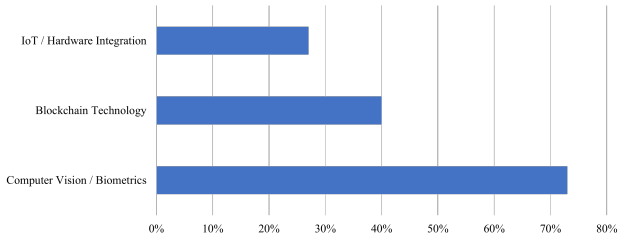


Figure 3. Frequency of major AI and NLP technology categories identified in the reviewed literature.

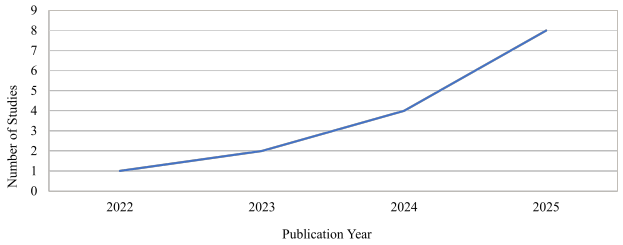


Figure 4. Publication trend of AI-powered short-answer grading studies across the review period.

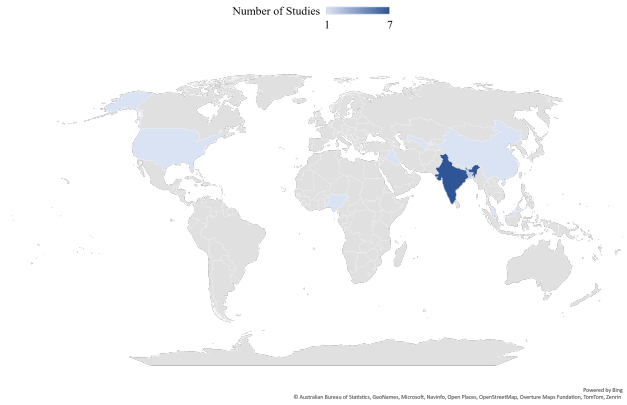


Figure 5. Geographical distribution of reviewed studies and related research activity.

4. DISCUSSION

The results of this Systematic Literature Review indicate a clear shift in automated short-answer grading research toward transformer-based models, which consistently demonstrate stronger performance than traditional machine learning approaches. This trend reflects the ability of contextualized language models to capture semantic nuances and linguistic variability in student responses. In contrast, earlier neural and similarity-based methods, including text similarity measures and question-answering-based frameworks, have shown competitive results primarily in controlled or domain-specific settings [8, 9, 13, 14, 11, 10].

A key limitation identified across the reviewed studies is the strong dependence of model performance on the availability of large, high-quality teacher-annotated datasets. This limitation is particularly pronounced for low-resource languages such as Uzbek, where annotated educational datasets are scarce or fragmented. As a result, many approaches reported in the literature struggle to generalize beyond specific domains or small-scale datasets, limiting their practical applicability in real educational environments.

The findings also suggest that multilingual and large language models offer promising directions for addressing data scarcity through transfer learning, few-shot learning, and zero-shot evaluation. However, challenges related to reproducibility, evaluation consistency, and transparency remain unresolved. Differences in dataset composition, grading scales, and evaluation metrics across studies further complicate direct comparison of reported results and hinder the establishment of standardized benchmarks.

Overall, the discussion highlights that while significant progress has been made in automated short-answer grading, the development of reliable systems for low-resource languages requires targeted dataset creation, careful model adaptation, and standardized evaluation frameworks. Future research should focus on constructing publicly available Uzbek short-answer datasets, exploring hybrid modeling approaches that combine linguistic knowledge with neural representations, and establishing consistent evaluation protocols to support fair and reproducible comparisons across studies.

5. ETHICS STATEMENT

This study adheres to established academic integrity and research ethics standards. The Systematic Literature Review was conducted exclusively using peer-reviewed and publicly available scholarly sources. No human participants, personal data, or proprietary datasets were involved in this research. All included studies were properly cited to avoid plagiarism, and the review process followed the PRISMA 2020 guidelines to ensure transparency, reproducibility, and methodological rigor. The findings reported in this paper are synthesized solely from existing literature, and no data were fabricated, manipulated, or experimentally generated by the authors.

6. CONCLUSION

This paper presented a Systematic Literature Review of AI-powered automated short-answer grading, with a particular focus on low-resource languages such as Uzbek. Following the PRISMA 2020 guidelines, 33 peer-reviewed studies published between 2018 and 2025 were systematically identified, analyzed, and synthesized to provide an overview of current research trends in this domain.

The review highlights the dominance of transformer-based models, which consistently outperform traditional machine learning approaches in automated short-answer grading tasks. Multilingual and large language models further demonstrate potential for addressing data scarcity through transfer learning, few-shot, and zero-shot evaluation. However, the findings also reveal that the limited availability of large, high-quality teacher-annotated datasets remains the primary barrier to deploying reliable automated grading systems in low-resource educational contexts.

Future research should prioritize the development of standardized and publicly available Uzbek short-answer datasets, the adaptation of transformer-based models to educational assessment tasks, and the establishment of consistent evaluation frameworks. Addressing these challenges is essential for building transparent, scalable, and reliable AI-powered grading systems that can support educational assessment in underrepresented languages.

REFERENCES

[1] S. Burrows, I. Gurevych, and B. Stein, "The eras and trends of automatic short answer grading," *International Journal of Artificial Intelligence in Education*, vol. 25, no. 1, pp. 60–117, 2015. doi: 10.1007/s40593-014-0026-8.

[2] M. Dzikovska, R. Nielsen, and C. Brew, "Automatic assessment of free text responses," in *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, 2013, pp. 1–9. [Online]. Available: https://aclanthology.org/W13-1701/

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[4] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.747/

[5] D. Alikaniotis, H. Yannakoudakis, and M. Rei, "Automatic text scoring using neural networks," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016. [Online]. Available: https://aclanthology.org/P16-1068/

[6] T. B. Brown et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, 2020. [Online]. Available: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[7] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021. doi: 10.1136/bmj.n71.

[8] B. Riordan, A. Horbach, A. Cahill, and T. Zesch, "Investigating neural architectures for short answer scoring," in *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, 2017. [Online]. Available: https://aclanthology.org/W17-5004/

[9] A. Horbach and T. Zesch, "A comparison of scoring short answers with human raters and automatic scoring," in *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, 2014. [Online]. Available: https://aclanthology.org/W14-1703/

[10] K. Taghipour and H. T. Ng, "A neural approach to automated essay scoring," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016. [Online]. Available: https://aclanthology.org/D16-1193/

[11] Z. Ke and H. T. Ng, "Question answering for automatic short answer grading," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. [Online]. Available: https://aclanthology.org/P19-1616/

[12] M. D. Shermis, "State-of-the-art automated essay scoring: Competition, results, and future directions," *Assessing Writing*, 2014. doi: 10.1016/j.asw.2013.04.001.

[13] A. Lauscher and T. Zesch, "A neural network model for automatic short answer grading," in *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, 2018. [Online]. Available: https://aclanthology.org/W18-0504/

[14] F. Dong and Y. Zhang, "Automatic short answer grading using text similarity," *Educational Technology & Society*, 2017. [Online]. Available: https://www.jstor.org/stable/90014594

[15] J. Cheng et al., "A survey of automatic short answer grading," *IEEE Transactions on Learning Technologies*, 2018. doi: 10.1109/TLT.2018.2852409.