-© 2026 Kumar et al. This article follows the Open Access policy of CC BY NC under CCA license v 4.0.

Submitted: 06/06/2026 - Accepted: 09/07/2026 - Published: 30/08/2026

Machine Learning-Enabled Automated Assessment and Grading in Education: A Systematic Literature Review of Techniques, Accuracy, and Fairness

Pritam Kumar[1*], Amarjeet Singh Mastana[2], Ximan Zong[3], Yunmei Wang[4]

[1]MSME Business School, Department of Digital Business Management, Assumption University, Bangkok, Thailand
[2]Faculty of Business Administration, St Teresa International University, Nakorn Nayok, Thailand
[3]MSME Business School, Department of Sustainable Business Management, Assumption University, Bangkok, Thailand
[4]MSME Business School, Department of Marketing, Assumption University, Bangkok, Thailand
*pkumar@msme.au.edu

DOI: 10.26417/9e7kn131

Abstract

This research aims to synthesize recent literature on machine learning-enabled automated assessment and grading in education, with emphasis on techniques, accuracy, fairness, and responsible implementation. A systematic literature review was conducted on 27 English-language peer-reviewed journal articles published between 2019 and 2025. The review focused on the major applications and evaluation issues of automated grading, including essay scoring, short-answer assessment, programming assessment, feedback generation, validity, explainability, and fairness. The findings show that natural language processing, transformer-based models, supervised machine learning, hybrid feature-based approaches, code execution, and testing frameworks are commonly used. Automated grading is more suitable for structured or semi-structured tasks, while complex writing, creativity, reasoning, and programming style require human moderation. Institutions should adopt automated grading as a supervised decision-support tool supported by validation, fairness auditing, transparency, appeal procedures, and teacher oversight. The paper contributes a focused synthesis linking techniques, accuracy, fairness, and governance in automated educational assessment.

Keywords: Automated feedback; Educational measurement; Automated Essay Scoring; Short-Answer Grading; AI-assisted Assessment; Rubrics based scoring; responsible AI

Introduction

Assessment is fundamental to education, yet manual marking is difficult to scale in large classes, online learning, and programming-intensive courses. Machine learning-enabled assessment has become important for scoring essays and short answers, evaluating programming assignments, generating feedback, and supporting decisions from learner performance data (Hussein et al., 2019; Rupp et al., 2019; Beseiso & Alzahrani, 2020). Recent studies apply deep learning, natural language processing, transformer models, programming-language analysis, and approaches to automate assessment (Xue et al., 2021; Uto, 2021; Faseeh et al., 2024; Misgna et al., 2024).

Automated assessment incorporates different methods depending on the task to be performed. For writing and short answer responses, automated assessment systems can assign grades based on the prediction of the score, semantic meaning analysis, linguistic feature analysis, or even provide formative feedback to students (Kumar & Boulanger, 2020; Beseiso et al., 2021; Zhang et al., 2019; Mizumoto & Eguchi, 2023; Hannah et al., 2023; Maslim et al., 2024; Ouahrani & Bennouar, 2024). However, automated assessment systems in programming education use test cases, code execution, static analysis, automated platforms, and feedback systems (Mekterović et al., 2020; Delgado-Pérez & Medina-Bulo, 2020; Combéfis, 2022). These systems can significantly reduce the workload for educators and provide near-instant feedback to students' assignments (Messer et al., 2024; Rojec et al., 2025; Mahdaoui et al., 2025). However, automated systems still require validation and supervision by educators when assigning grades and supervising student learning.

Due to concerns regarding accuracy, reliability, validity, fairness, explainability, transparency, and accountability of AI systems, there is a need for evaluating these models based on aspects beyond their accuracy in predicting the outcome of specific tasks and adapting to particular languages. Measures of agreement between human and machine evaluators of student work are of particular interest in the literature, as are aspects of robustness, suitability of tasks for automation, quality of feedback, and fairness in the evaluation of student groups (Ramesh & Sanampudi, 2021; Conijn et al., 2023; Andersen et al., 2025; Alqurashi et al., 2025; Lohmann et al., 2025).

Problem Statement

Automated assessment using machine learning technologies is increasingly being used to assess student written assignments, such as essays, short answers, programming assignments, and even feedback submitted by students to teaching systems. Although automated assessment technologies can reduce the workload of teachers and enable the faster obtaining of assessment results from submitted assignments, there are concerns regarding the use of these technologies in education,

such as inaccuracies in the scoring of student assignments, lack of validity of the automated scoring systems, lack of fairness in automated assessment technologies, lack of transparency in the automated assessment technologies, and the lack of accountability of these technologies. Many existing studies on automated assessment technologies focus on the technical aspects of the technologies, but pay less attention to educational aspects, such as consequences of using automated assessment technologies, diversity of students, human oversight of automated assessment technologies, and quality of feedback provided to students by automated assessment technologies. Furthermore, automated assessment technologies are generally less capable of evaluating aspects of student assignments related to creativity, critical thinking, complex thinking and reasoning, writing quality, and programming style. Consequently, there is a need for a review of the literature regarding automated assessment technologies to determine the different techniques used in these technologies, the accuracy of these technologies, the fairness of these technologies, and the requirements for implementing these technologies.

Research Objectives

The general objective of this study is to synthesize recent literature on machine learning-enabled automated assessment and grading, with particular attention to the techniques used, reported accuracy considerations, and fairness implications.

The specific objectives are as follows:

RO1: To identify the main machine learning techniques and assessment contexts reported in recent research on automated assessment and grading in education.

RO2: To examine how the reviewed literature discusses accuracy, validity, feedback quality, and task suitability in automated grading systems.

RO3: To analyze the extent to which fairness, transparency, bias, and ethical implementation are addressed in studies of machine learning-enabled assessment.

RO4: To identify the methodological, technical, educational, and governance limitations that restrict the responsible adoption of machine learning-enabled automated assessment systems in educational institutions.

RO5: To identify research gaps and propose directions for future empirical studies on automated grading accuracy and fairness.

Research Questions

This review is guided by the following research questions:

RQ1: What machine learning techniques and educational assessment contexts are most commonly associated with automated assessment and grading in the reviewed literature?

RQ2: How does the literature evaluate or discuss the accuracy, reliability, validity, and task suitability of machine learning-enabled automated grading systems?

RQ3: What fairness-related concerns, including bias, transparency, explainability, and equitable treatment of learners, are identified or implied in the reviewed studies?

RQ4: What limitations in the current literature restrict the responsible adoption of automated assessment in educational institutions?

RQ5: What future research directions are needed to strengthen the accuracy, fairness, transparency, and educational usefulness of machine learning-enabled grading systems?

Research Gap

There are several research gaps in the literature that justify the undertaking of this review. For instance, research on automated assessment of written and coded assignments and feedback systems exists in areas like essay assessment, short answer assessments, programming assessments, fairness of automated systems, and validity of those automated systems (Hussein et al., 2019; Ramesh & Sanampudi, 2021; Deepshikha, 2025). In addition, many studies report on the accuracy of automated systems to grade written and coded assignments, but provide little evidence in relation to the accuracy of those automated systems in comparison to expert graders of the assignments (Xue et al., 2021; Hannah et al., 2023; Maslim et al., 2024; Lohmann et al., 2025). Furthermore, there is a lack of research publication regarding the fairness and transparency of automated assessment systems; it is possible for these types of systems to disadvantage certain learners relative to others based upon their language background, writing styles, coding approaches, and more (Andersen et al., 2025; Alqurashi et al., 2025). Additionally, much of the research that exists regarding the automated assessment of programming assignments reveals the potential for such systems to greatly facilitate the grading of programming assignments, but mostly focuses upon the correct functionality of the submitted programs rather than the readability, maintainability, and overall quality of those programs (Delgado-Pérez & Medina-Bulo, 2020; Combéfis, 2022; Rojec et al., 2025; Mahdaoui et al., 2025). Finally, there exists a research gap in relation to studies that investigate the role of teachers in automated grading systems, the usefulness of feedback from automated systems, the role of student trust in those systems, and any related governance of those automated systems.

Conceptual and Educational Assessment Background

Educational assessment involves collecting and interpreting evidence about student learning in relation to intended learning outcomes. In automated grading, high model accuracy alone is insufficient unless the score meaning is educationally valid and aligned with the intended construct, rubric, and assessment purpose (Barredo Arrieta et al., 2019). Validity in machine learning-enabled assessment requires that the

system measures the relevant learning outcome. For example, for essay scoring, the system should assess writing quality, and for programming assessment, it should consider the quality of the code produced by the student (Zhai et al., 2020; Ramesh & Sanampudi, 2021).

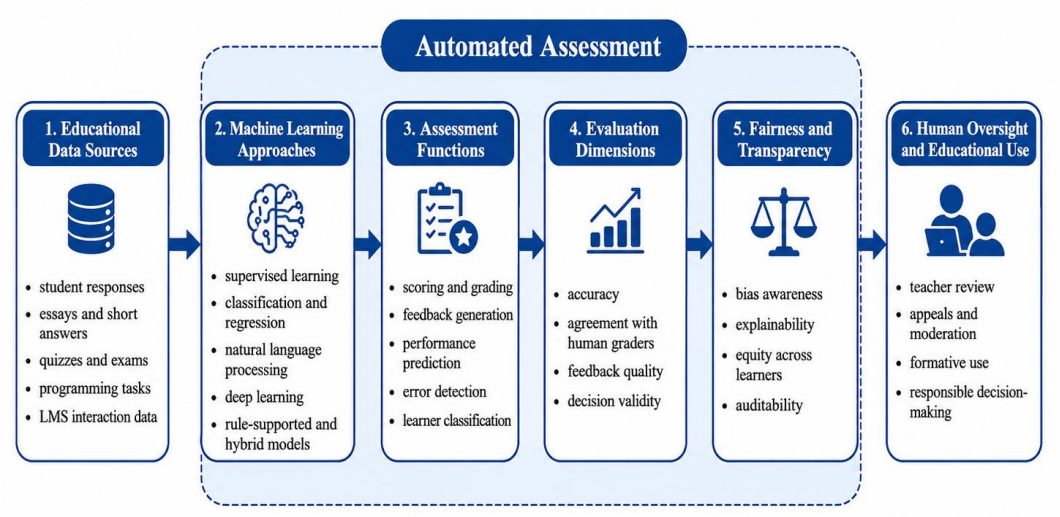
Reliability of automated assessment refers to the consistency and defensibility of the assessment scores obtained from various tasks, prompts, learners, and assessment contexts. Such attributes include the agreement between humans and machines, the stability of the machines, and how robust the automated assessment is across different datasets and educational contexts (Conijn et al., 2023). Automated assessment systems may be applied to formative assessment processes, where feedback from automated systems optimizes the learning process for each learner. However, the use of automated systems for summative assessments requires stronger validity and fairness properties (Messer et al., 2024). Automated systems must exhibit properties of fairness to exclude any form of bias during the automated assessment process (Mehrabi et al., 2021; Alqurashi et al., 2025). Therefore, the focus of this research is on the educational assessment lens through which the properties of automated assessment using machine learning can be examined. The focus will be on properties like validity, reliability, feedback quality, fairness, explainability, transparency, and human oversight.

Machine learning-enabled automated assessment involves using algorithms to assess student work and determine scores. Furthermore, automated assessment includes several areas of study, including automated essay scoring, short-answer grading, programming assessment, automated feedback, explainability, validity, and fairness. These areas demonstrate that machine learning-enabled automated assessment goes beyond simply determining a student's score on an assignment but also involves feedback for the assignment, suitability of the assessment task for the student, and other factors related to the automated assessment of student work (Hussein et al., 2019; Ramesh & Sanampudi, 2021; Messer et al., 2024; Misgna et al., 2024). In the area of assessing written responses, there are a variety of different approaches that can be taken by machine learning systems, such as utilizing features of the language, embeddings, similarities in meaning, neural models, or even a combination of these approaches (Beseiso & Alzahrani, 2020; Xue et al., 2021; Maslim et al., 2024; Lohmann et al., 2025). Within the area of assessing programming, different approaches can be taken to automatically assess student submissions, such as through the use of test cases (Mekterović et al., 2020; Combéfis, 2022; Messer et al., 2024; Rojec et al., 2025; Cisneros-González et al., 2025).

As is illustrated in Figure 1, the conceptual background of the present research relates the data sources to the machine learning approaches, the assessment and function aspects, the evaluation dimensions, fairness, and human oversight. Since the automated grading of assignments can have an impact upon fairness, bias, transparency, robustness, and the need for human oversight of the grading process

(Barredo Arrieta et al., 2019; Mehrabi et al., 2021; Andersen et al., 2025), it is essential for the machine learning approach to support the automated grading of assignments, but for the teachers to still have oversight and responsibility for the fairness of such automated grading of assignments.

Figure 1. Conceptual background of machine learning-enabled automated assessment and grading



Source: Developed by the author based on the reviewed literature.

Methodology

Review Design

To examine the topic of machine learning-enabled automated assessment and grading in education, a PRISMA-guided systematic literature review was employed. The PRISMA approach ensures that the process of finding, screening, qualifying, and including the studies in the review is systematic and transparent (Page et al., 2021). A systematic literature review was used for this study because the literature review is tasked with reviewing several different types of tasks that have been performed using automated grading systems.

Information Sources and Search Strategy

The research databases that were searched in the literature review included Google Scholar, Scopus, Web of Science, IEEE Xplore, SpringerLink, and the ACM Digital Library. Publish or Perish 8 was used to retrieve, organize, and export articles from Google Scholar, Scopus, and Web of Science. Additionally, manual searches were also performed in IEEE Xplore, SpringerLink, the ACM Digital Library, the publisher webpages of the articles of interest, the DOI (digital object identifier) of the articles of interest, and the websites that published the articles. Each search utilized the

following keywords and limits: machine learning, artificial intelligence, deep learning, natural language processing, transformer, BERT, automated assessment, automated grading, automatic scoring, automated essay scoring, short-answer grading, programming assessment, code grading, education, higher education, classroom, student, accuracy, validity, feedback, fairness, bias, explainability, transparency. The date limits for the articles were published between 2019 and 2025.

The main search string for the literature review was as follows: ("machine learning" OR "artificial intelligence" OR "deep learning" OR "natural language processing" OR transformer OR BERT) AND ("automated assessment" OR "automated grading" OR "automatic scoring" OR "automated essay scoring" OR "short-answer grading" OR "programming assessment" OR "code grading") AND (education OR "higher education" OR classroom OR student) AND (accuracy OR validity OR feedback OR fairness OR bias OR explainability OR transparency). Additionally, narrow searches were performed on topics relating to automated essay scoring fairness, short-answer grading, programming assessment feedback, AI grading validity, and explainable automated scoring. Through these search strategies, 1,940 articles were found through the literature search.

Eligibility Criteria

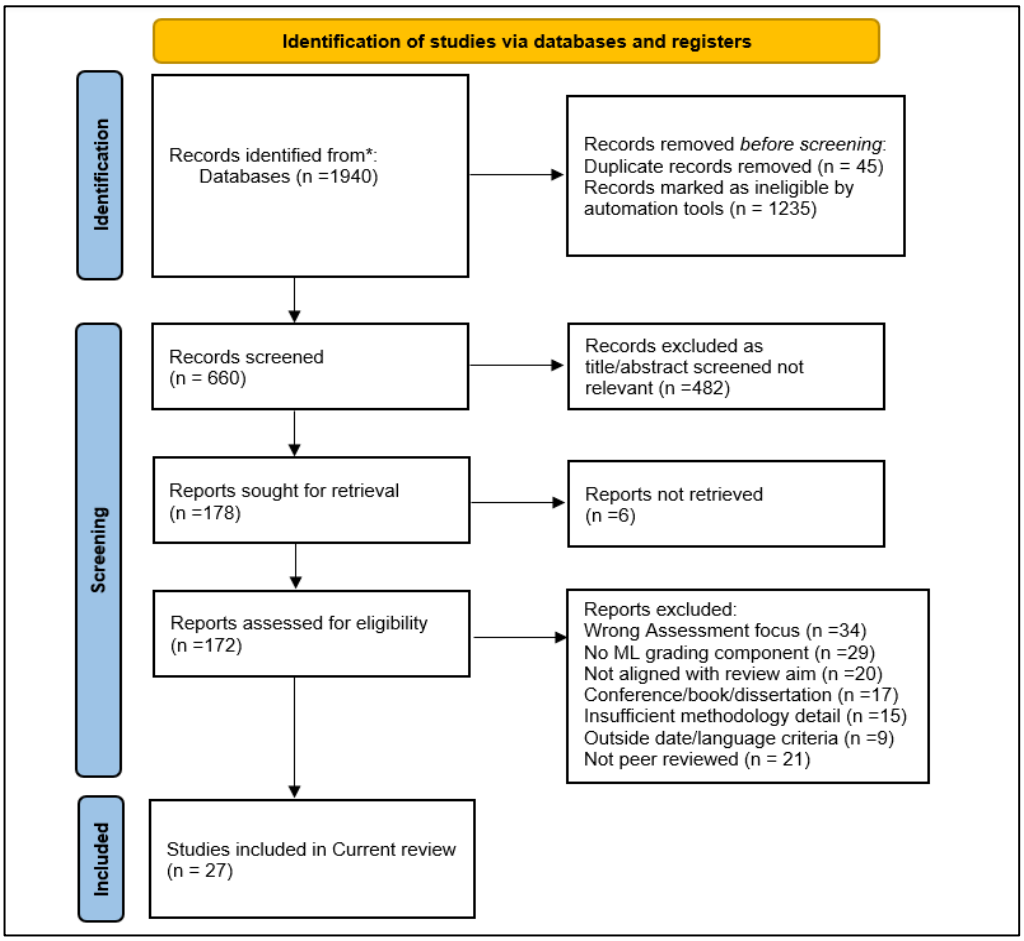
Studies were included when they were published between 2019 and 2025, written in English, and directly related to machine learning-enabled automated assessment or grading. Eligible studies were peer-reviewed journal articles, including systematic review articles, empirical model-development studies, system-evaluation studies, and assessment-validity studies where they directly addressed automated essay scoring, short-answer grading, programming assessment, automated feedback, validity, fairness, transparency, or explainability. Conference papers, book chapters, dissertations, technical reports, non-journal publications, and non-peer-reviewed outputs were excluded. Studies were also excluded when they did not address automated assessment or grading, were unrelated to education, lacked sufficient methodological detail, or focused only on general AI in education without an assessment component.

Screening and Study Selection

As shown in Figure 2, the initial search identified 1,940 records. Before title and abstract screening, 45 duplicates were removed. After screening the databases for relevance, 1,235 articles from supplementary searches were eliminated, leaving 660 records to be screened for title and abstract screening. Of the 660 screened records, 482 were excluded because they did not meet the inclusion criteria. In total, 178 reports were sought for retrieval, of which 6 were not retrieved. The remaining 172 full-text reports were assessed for eligibility. At this stage, 145 reports were excluded for reasons including wrong assessment focus, absence of a machine learning grading component, insufficient methodological detail, date or language mismatch, non-

journal publication type, or non-peer-reviewed status. Therefore, 27 peer-reviewed journal articles were included in the final review.

Figure 2. PRISMA 2020 flow diagram for study selection in the current review (n = 27)



Source: Adapted from Page et al. (2021).

Data Extraction

Data extraction in the review included author names, citation counts, publication year, region, research objectives, key findings, research type, research design, sample size, data analysis method, and machine learning applications or techniques. For the present article, these extracted data have been reorganized to emphasize the topic of automated assessment and grading. Particular attention has been given to each of the following data elements: the type of assessment task, the machine learning or AI technique, the use of learning analytics or predictive analytics, the feedback or

grading assigned, the benefits of such automated systems, and the limitations to such automated systems.

Data Synthesis and Coding

The synthesis combined content analysis and thematic analysis. Content analysis was used to summarize publication year, geographical scope, research type, study design, evidence source, data analysis method, and application area. Thematic analysis was used to interpret how automated assessment was described across the reviewed peer-reviewed journal articles. Coding focused on five overlapping thematic categories: (1) automated essay and writing scoring; (2) short-answer and text-response grading; (3) programming and code assessment; (4) feedback, explainability, trust, and human oversight; and (5) validity, robustness, fairness, and bias. For each category, the review also recorded the principal automated technique, intended assessment use, reported strength, and main limitation. The resulting application-area distribution and task-based technique synthesis are presented in the Results section.

Quality, Accuracy, and Fairness Lens

The included studies were evaluated through a Quality-Accuracy-Fairness (QAF) lens because the body of reviewed studies comprised diverse designs, including model-development, system-evaluation, review, validity, and fairness studies. A single risk-of-bias instrument was therefore unsuitable. Instead, appraisal considered the clarity of each study's aim, research design, educational context, data source, model description, evaluation procedure, and relevance to automated assessment (Hong et al., 2018; Page et al., 2021). The quality dimension examined whether methods, datasets, systems, and findings were reported with sufficient detail to determine whether the research questions were appropriately addressed. Review studies were assessed for transparent search and selection procedures, whereas empirical and system-development studies were examined for clear dataset descriptions, model specifications, and evaluation methods. The accuracy dimension considered technical performance metrics, agreement between automated scores and judgments from teachers or experts, robustness, and the suitability of particular assessment tasks. Accuracy was interpreted as both the ability of the model to make correct predictions about student performance, as well as the defensibility of the educational judgments that must be made about the students based upon the model's predictions (Zhai et al., 2020; Ramesh & Sanampudi, 2021). The dimension of fairness was examined to determine if the models exhibited bias, error rates differed among subgroups of students, were transparent and explainable in their predictions of student performance, and if the models exhibited human oversight in making their predictions (Barredo Arrieta et al., 2019; Mehrabi et al., 2021). Overall, the framework guided the development of the narrative of the review study, rather than leading to the calculation of a numerical score based upon the included studies. Table 1

summarizes the criteria that were used to apply the evaluation to the reviewed studies.

Table 1. Quality-Accuracy-Fairness (QAF) appraisal method for automated grading

QAF element	Purpose	Recommended practice	Method/source
Study quality	Checks whether the evidence is methodologically credible	Assess clarity of aims, design, data source, review procedure, and relevance to education	PRISMA reporting and MMAT-style appraisal (Page et al., 2021; Hong et al., 2018)
Human-grade comparison	Checks alignment with expert judgment	Compare machine scores with teacher scores and report agreement	Agreement validation, including kappa for categorical scores where applicable (Cohen, 1960)
External validation	Tests generalizability	Evaluate models on new cohorts, courses, or institutions	Validation logic and model documentation across intended-use contexts (Barredo Arrieta et al., 2019)
Subgroup fairness analysis	Detects unequal error patterns	Check performance across relevant learner groups while protecting privacy	Fairness auditing using subgroup error rates, bias analysis, and differential performance checks (Mehrabi et al., 2021)
Explainability and documentation	Supports trust, feedback, and accountability	Provide reasons for scores, model limitations, and appropriate-use conditions	Explainable AI and transparent reporting principles (Barredo Arrieta et al., 2019)
Uncertainty handling	Reduces unfair grading of difficult cases	Flag low-confidence submissions for human review	Human-in-the-loop moderation and escalation of uncertain cases
Appeal and moderation	Protects student rights	Allow teachers and students to challenge or review automated grades	Assessment governance with teacher oversight and audit trails

Results

The results are presented in a clear manner that summarize what was found in each of the 27 publications. The results describe a discussion of publication trends, author locations, methods of publications, data sources, analysis methods used by authors, applications of automated assessment and grading, automated assessment techniques found in the publications, accuracy of automated assessment techniques, and an overall summary of studies.

Publication Trend

The publication trend is based on the journals that were reviewed, which were published between 2019 and 2025. Older articles focused primarily on essay scoring, short answer questions, and programming assignments. However, recent articles talk about explainability, the use of feedback from the systems, validity, fairness of the models, trustworthy scoring systems, and ethics of AI. Table 2 shows the number of publications in this field over the years, indicating growth in 2025 compared to previous years.

Table 2. Publication trend of reviewed studies, 2019-2025

Year	No. of studies	Percentage	Interpretation
2019	3	11.1%	Foundational work on automated essay scoring, large-scale scoring, and short-answer grading.
2020	4	14.8%	Research on BERT-based essay scoring, explainable rubric scoring, programming assessment, and automated assessment systems.
2021	4	14.8%	Studies on reliable essay scoring, automated essay-scoring reviews, deep-neural models, and multi-dimensional scoring.
2022	1	3.7%	A review of automated code-assessment techniques, tools, classifications, and future directions.
2023	3	11.1%	Research on scoring validity, student trust in explanations, and AI-supported essay scoring.
2024	5	18.5%	Recent work on hybrid essay scoring, trustworthy short-answer grading, programming autograding, feedback, and explainability.
2025	7	26.0%	Current research on AI-enabled grading, fairness, programming assessment, large language models, and responsible implementation.
Total	27	100.0%	All 27 peer-reviewed journal articles included in the review.

Regional Distribution

The studies reviewed in this research originate from various scholarly contexts. Since some of the studies are based upon reviews of existing literature or the implementation of models based upon existing datasets, it is inherently difficult to draw conclusions regarding the regional distribution of these studies. Therefore, Table 3 represents the geographical scope of the reviewed studies as a whole rather than reviewing each study as if it represented the entirety of a given nation's study on automated grading systems.

Table 3. Regional distribution of reviewed studies

Geographical scope	No. of studies	Percentage	Relevance to automated assessment
International or review-based scope	6	22.2%	Synthesizes evidence across multiple countries, tasks, and assessment systems.
North America	4	14.8%	Important for AES, student trust, and automated scoring validity studies.
Europe	7	25.9%	Strong representation in AES, programming assessment, feedback, and explainability studies.
Asia and Middle East	9	33.4%	Relevant to BERT-based scoring, multilingual short-answer grading, and programming assessment.
Other or not clearly specified	1	3.7%	Included where the study context was technical or cross-dataset rather than geographically specific.
Total	27	100.0%	Regional classification should be interpreted cautiously because some studies are reviews or dataset-based.

Methodological Characteristics

The methodological profiles of the studies presented in Tables 4–7 is technically diverse. The studies include model development studies, system-evaluation studies, review studies, fairness and validity studies, and user-centered studies regarding feedback and trust. Thus, while these studies provide some means of performing a synthesis of the field, the results of these studies should be understood as a review of the field rather than a statistical analysis of the results of comparable studies.

Table 4. Research type of reviewed studies

Research type	No. of studies	Percentage	Implication
Empirical/model-development studies	14	51.9%	Dominant evidence types for testing scoring models and automated assessment methods.
System or tool implementation studies	5	18.5%	Shows practical grading workflows and platform-level implementation.
Review or survey studies	5	18.5%	Provides broader synthesis of automated scoring and programming assessment.
Validity, fairness, explainability, or robustness studies	3	11.1%	Highlights responsible use beyond accuracy alone.
Total	27	100.0%	The reviewed studies are technically diverse.

Table 5. Research design of reviewed studies

Research design	No. of studies	Percentage	Implication for review rigor
Experimental/modeling design	12	44.5%	Tests automated scoring or grading models on learner responses or code.
System design and evaluation	5	18.5%	Examines auto grading infrastructure, workflows, or implementation.
Literature review or survey	5	18.5%	Synthesizes existing evidence and identifies trends.
Dataset, benchmark, validity, or fairness analysis	3	11.1%	Examines validity, bias, robustness, or evaluation quality.
User, trust, or feedback-utilization study	2	7.4%	Addresses educational use, student response, or explanation value.
Total	27	100.0%	Multiple designs support thematic synthesis.

Table 6. Evidence sources and sample reporting in the reviewed studies

Evidence source	No. of studies	Percentage	Interpretation
Open datasets or text/code corpora	10	37.1%	Common in AES, short-answer scoring, and NLP model testing.
Classroom or institutional submissions	7	25.9%	Supports ecological validity for programming and writing assessment.
System, toolkit, or platform evaluation	4	14.8%	Useful for implementation and workflow evidence.
Literature as data	5	18.5%	Review studies synthesize prior evidence instead of primary learner data.
Mixed or secondary evidence	1	3.7%	Uses broader conceptual or methodological evidence related to assessment systems.
Total	27	100.0%	Evidence sources are heterogeneous across task types.

Table 7. Main data analysis or evaluation approaches used in the reviewed studies

Main analysis/evaluation approach	No. of studies	Percentage	Interpretation
NLP, deep learning, or transformer modeling	10	37.1%	Most common technical approach for written responses and semantic scoring.
Traditional or hybrid ML with handcrafted features	5	18.5%	Combines statistical learning with linguistic, rubric-based, or feature-based evidence.
System, testing, or workflow evaluation	5	18.5%	Common in programming auto grading and feedback tools.
Review or thematic synthesis	5	18.5%	Used to summarize existing automated assessment research.
Fairness, explainability, validity, or robustness analysis	2	7.4%	Addresses responsible use and trustworthiness.
Total	27	100.0%	Approaches align with the reviewed journal articles.

Machine Learning Application Areas

Table 8 presents an overview of the main applications of each of the reviewed studies. The main areas of application within the studied technologies and methodologies include automated essay scoring, short-answer question grading, programming and code grading, automated feedback technologies, explainability technologies, fairness technologies, validity technologies, and robustness technologies. Some studies included in this review contributed to more than one of these categories.

Table 8. Major automated assessment application areas in the reviewed studies

Application area	No. of studies	Percentage	Meaning for automated assessment
Automated essay/writing scoring	12	44%	Supports scoring and feedback for extended written responses.
Short-answer/text-response grading	6	22%	Supports concept-level scoring and rapid formative assessment.
Programming/code assessment	7	26%	Supports auto grading, test-case evaluation, and programming feedback.
Automated feedback/explainability/trust	8	30%	Links scoring outputs to student and teacher understanding.
Fairness/validity/robustness	7	26%	Examines whether automated scores are defensible, equitable, and reliable.
Note	Categories are not mutually exclusive.	Percentages use n = 27.	

Automated Assessment Techniques and Task Suitability

The reviewed evidence indicates that the suitability of an automated-assessment technique depends primarily on the structure of the task, the clarity of the scoring criteria, and the educational consequences of the resulting grade. Table 9 summarizes the relationship between assessment tasks, commonly reported techniques, likely strengths, and principal limitations. The table represents a thematic synthesis of the reviewed studies rather than a universal classification of all automated-assessment systems.

Table 9. Automated assessment techniques by task type

Assessment task type	Common ML or automated technique	Likely strength	Main limitation
Essays and extended writing	AES models, transformers, hybrid features, rubric prediction	Useful for preliminary scoring and feedback at scale	Risk of weak construct validity, prompt sensitivity, and language bias
Short factual or conceptual responses	Semantic similarity, SBERT/BERT, deep learning, classification	Useful for rapid formative feedback and concept checking	May miss alternative correct expressions or nuanced reasoning
Programming assignments and exams	Code execution, test cases, static analysis, CodeBERT, autograder workflows	Effective for correctness and rapid feedback	Weak for readability, maintainability, documentation, and design quality
Automated feedback systems	NLP feedback generation, explanation interfaces, error diagnosis	Can support revision and student self-regulation	Feedback may be generic, unclear, or over-trusted
Fairness and validity evaluation	Robustness testing, subgroup analysis, validity arguments, explainability	Improves accountability and trustworthiness	Often reported inconsistently or after model development
High-stakes or complex tasks	Semi-automated grading and human moderation	Combines machine efficiency with teacher judgment	Requires governance, appeals, and instructor workload planning

Accuracy Findings

The accuracy-related synthesis within Table 10 indicates that accuracy is strongest with the application of automation to tasks that have expected outputs, scoring rules, or rubrics. However, the reviews of these journal articles also indicate that accuracy is more than just the performance of the predictions alone. Accuracy in the automated grading context should also relate to the defensibility of the automated grades, the agreement between the automated and human grades, the robustness of the automation system, and the validation of the automated grades relative to some external standard.

Table 10. Accuracy-related findings for automated grading

Accuracy dimension	What it means	Evidence from synthesis	Required improvement
Human-grade agreement	Similarity between machine scores and expert human scores	Central for AES and short-answer systems but not always reported in comparable form	Report human-machine agreement and inter-rater reliability
Task suitability	Match between model capability and assessment task	Stronger for constrained answers and functional programming tests than creative or interpretive tasks	Match automation level to assessment purpose
Robustness	Performance on new prompts, cohorts, unusual answers, or adversarial responses	Robustness concerns appear in AES and scoring validity research	Test across prompts, datasets, institutions, and response styles
Feedback validity	Usefulness and correctness of automated feedback	Feedback systems show promise but require learning-impact evaluation	Evaluate whether feedback improves revision and learning outcomes
External validation	Generalization beyond original dataset or course	Often limited in model-development studies	Use independent datasets and classroom replication

Fairness, Transparency, and Human Oversight

As summarized in Table 11, fairness received considerably less direct attention in the 27 included automated-assessment studies than efficiency and accuracy. Within this evidence base, only some studies explicitly examined algorithmic fairness, subgroup performance, language-related bias, transparency, or explainability in automated scoring systems (Conijn et al., 2023; Maslim et al., 2024; Andersen et al., 2025; Alqurashi et al., 2025). These included-study findings therefore support the conclusion that fairness evaluation remains uneven and insufficiently standardized in automated assessment research. Broader concerns about training-data bias, differential outcomes across learner groups, and the need for transparent and explainable systems are additionally informed by general AI fairness and explainability literature rather than by all 27 included automated-assessment studies (Barredo Arrieta et al., 2019; Mehrabi et al., 2021). Accordingly, contextual factors such as language background, educational setting, curriculum, and learner characteristics should be treated as priorities for future empirical testing rather than as effects conclusively demonstrated across the complete review sample.

Table 11. Fairness and transparency findings

Issue	Risk in automated assessment	Result from synthesis	Recommended response
Training-data bias	Past grading patterns or dataset imbalance may be reproduced	Fairness is addressed directly in only some reviewed studies.	Audit errors across relevant learner groups and prompts
Language and expression bias	Students may be penalized for writing style, dialect, or language background	Especially relevant to AES and short-answer scoring	Use diverse data, bias testing, and human review
Black-box scoring	Students and teachers may not understand grades	Explainability and model documentation remain uneven	Provide score reasons, limitations, and appeal pathways
Programming-style bias	Correct code may be penalized if it differs from expected patterns	Auto graders can overemphasize functional tests	Assess readability, documentation, and design with teacher moderation
Over-automation	Teachers may lose professional judgment in high-stakes decisions	Semi-automated systems appear more responsible	Keep teachers in control of final grades and uncertain cases

Discussion

The following discussion interprets the detailed results in relation to the five research questions, with particular attention to technique suitability, assessment quality, fairness, responsible adoption, and priorities for future research.

9.1 To address RQ1, the findings indicate that automated assessment using machine learning is not a one-size-fits-all solution but relies upon different approaches to assessment according to the type of assessment that is to be performed. Assessments of essays and writing assignments indicate a shift towards the use of neural networks, transformers, hybrid linguistic features models, and NLP models, though each of these approaches still require validation to ensure that the written work submissions meet the requirements of the writing assignment (Hussein et al., 2019; Beseiso & Alzahrani, 2020; Xue et al., 2021; Faseeh et al., 2024). Short answer questions can be assessed for semantic meanings of the answers with explainable and multilingual AI models (Zhang et al., 2019; Maslim et al., 2024; Alqurashi et al., 2025). Finally, programming assignments can be automatically assessed due to the suitability of automated systems to assess code for correctness (Mekterović et al., 2020; Combéfis, 2022; Rojec et al., 2025). Thus, automated assessment using machine learning models can be

beneficial for different types of assignments according to their inherent suitability for automation.

In addressing RQ2, the reviewed evidence indicates that automated grading is most accurate when assignments are structured around clear and explicit scoring expectations. Auto grading systems are able to grade programming assignments. Short answer questions can be graded if the concepts that are to be answered are understood. Essay questions can be automatically graded, although it is more difficult to obtain the validity of such automatic grading systems. While the accuracy of auto grading systems is important, studies in educational journals indicate that accuracy is not enough for automatic grading systems to be considered valid. Studies show that the automated grades must also correspond with rubrics, expert judgment, and tests for robustness (Kumar & Boulanger, 2020; Ramesh & Sanampudi, 2021; Hannah et al., 2023; Maslim et al., 2024; Lohmann et al., 2025). Thus, accuracy in automatic grading is not just a measurement of the AI model's performance, but also of the validity of the automatic grading system.

With regard to RQ3, the included automated-assessment studies addressed fairness, transparency, and explainability unevenly. Direct evidence within the review was concentrated in a limited group of studies examining explanations, trustworthiness, language-related scoring, and algorithmic fairness in essay and short-answer assessment (Conijn et al., 2023; Maslim et al., 2024; Andersen et al., 2025; Alqurashi et al., 2025). These studies indicate that automated grading systems should communicate their limitations, provide understandable reasons for assigned scores, and test whether performance differs across relevant learner groups. However, claims concerning broader risks—such as the reproduction of historical bias, systematic disadvantage across demographic groups, and general requirements for subgroup fairness—draw partly on wider explainable-AI and machine-learning fairness literature rather than solely on the 27 included automated-assessment studies (Barredo Arrieta et al., 2019; Mehrabi et al., 2021). The QAF lens therefore uses these broader principles as an interpretive framework, while the review findings themselves support the more limited conclusion that fairness testing, transparency, and subgroup reporting remain inconsistently addressed in the included evidence base.

In relation to RQ4, the synthesis identifies the heterogeneity of the existing studies as a major limitation affecting responsible institutional adoption. Some studies are empirical model development studies while others are review articles. Another limitation is that many of the studies include information about the technical systems but less information about human-grade evaluation of the systems. In general, programming systems are more effective at evaluating the correctness of code than the readability or maintainability of the code (Mekterović et al., 2020; Delgado-Pérez & Medina-Bulo, 2020; Combéfis, 2022; Messer et al., 2024). Writing and short-answer question systems often have a higher degree of difficulty in finding evaluation studies

that provide evidence of their robustness, lack of bias, and the validity of the feedback that they provide to students (Kumar & Boulanger, 2020; Ramesh & Sanampudi, 2021; Hannah et al., 2023; Maslim et al., 2024; Conijn et al., 2023).

In response to RQ5, the reviewed evidence highlights the need for a coordinated research agenda that strengthens the accuracy, fairness, transparency, contextual relevance, and educational usefulness of machine learning-enabled grading systems. Accuracy studies should use measures that go beyond using a single dataset to test the model's predictions against expert graders' predictions, testing the reliability of the models across various metrics and the robustness of the models to unusual answers from students (Conijn et al., 2023; Maslim et al., 2024; Lohmann et al., 2025). The accuracy of automated grading systems can also be measured by comparing the answers of students to various types of assessments, such as essays, short answers, programming assignments, and formative and summative assessments (Ramesh & Sanampudi, 2021; Messer et al., 2024; Rojec et al., 2025; Mahdaoui et al., 2025).

The research on fairness should evaluate the performance of these models across different aspects of their populations, such as language background, gender, disability, educational level, writing style, and coding approach (Mehrabi et al., 2021; Andersen et al., 2025; Alqurashi et al., 2025). Furthermore, the models developed in one institution or culture might not perform equally well in another, thus requiring additional studies across several institutions and cultures to ensure that the models are performing fairly within the populations of those institutions (Messer et al., 2024; Alqurashi et al., 2025). Additional steps can be taken to make the models more transparent to the end users, such as revealing the reasoning behind the scores given to submissions, providing feedback, audits, model documentation, and making appeals available to the users (Barredo Arrieta et al., 2019; Deepshikha, 2025).

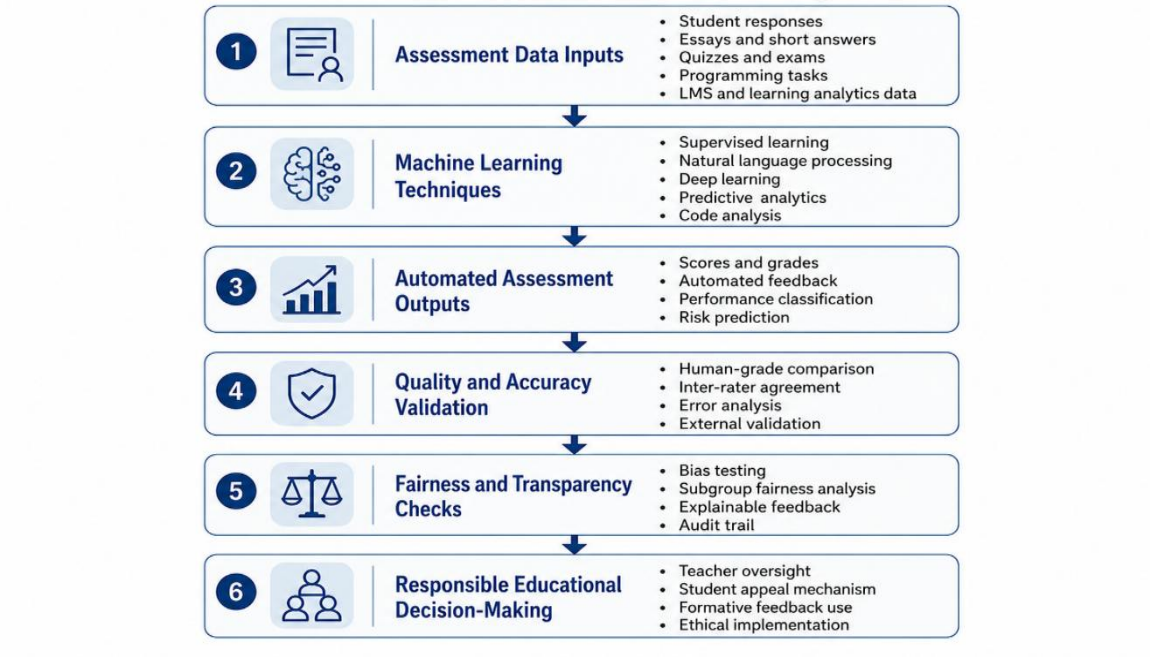
The educational usefulness of automated grading systems can only be ascertained through longitudinal and classroom-based studies into the learning of students who used these systems (Zhang et al., 2019; Conijn et al., 2023). Automated grading systems should be designed to assist teachers, not to replace them. Studies that incorporate teachers into the grading process can determine in what situations teachers should review automated grades and how this process will impact both the grading of assignments by teachers, and the confidence that students have in those grades. Therefore, teacher training should include these topics. Furthermore, for assignments that utilize rubrics, automated grading systems can be designed to reflect these rubrics, while still allowing for teachers to retain the ability to grade assignments for components like originality, critical thinking, creativity, and professional judgment. Finally, a governance framework can be established to determine how automated grading systems should be regulated before they are applied to grading for various disciplines and levels of education.

Proposed Framework for Responsible Machine Learning-Enabled Assessment

Figure 3 presents a framework for automated assessment and grading using machine learning. The framework receives assessment data as inputs from the students, such as the answers that the students provide for the short answer questions, essay questions, quizzes, exams, and data from the learning management systems. The quality of this data must be ensured before the automated assessment system can utilize this data to learn from it (Hannah et al., 2023; Lohmann et al., 2025; Mehrabi et al., 2021). Machine learning techniques are applied to the assessment data, such as supervised machine learning, natural language processing, deep learning, predictive analytics, and code analysis techniques. These techniques should be chosen according to the intention of the automated assessment system.

For instance, automated systems can assess structured or semi-structured tasks, including rubric-based writing and programming tasks with clear functional requirements, whereas open-ended writing, creativity, complex reasoning, programming architecture, maintainability, and style require stronger human evaluation (Beseiso et al., 2021; Xue et al., 2021; Combéfis, 2022; Mahdaoui et al., 2025). The automated system outputs scores for each assessment task. These scores should be checked by humans to ensure the accuracy of the automated system's scoring techniques (Rupp et al., 2019; Hannah et al., 2023; Maslim et al., 2024). The system should also be checked for fairness and transparency in its automated grading process. For instance, the automated system could be checked for biases in its programming and the fairness of its outcomes for different students using techniques such as bias testing, fairness testing, explainable machine learning, and audit trails (Barredo Arrieta et al., 2019; Mehrabi et al., 2021; Andersen et al., 2025; Alqurashi et al., 2025). Finally, a teacher should be involved in the automated assessment and grading system to ensure ethical implementation of the automated grading system (Conijn et al., 2023; Messer et al., 2024; Deepshikha, 2025).

Figure 3. Proposed Framework for Responsible Machine Learning-Enabled Automated Assessment and Grading



Source: Developed by the author based on the synthesis of the 27 reviewed journal articles and the QAF appraisal lens.

Implications

The findings of this research have implications for various groups, including educators, educational institutions, system developers, researchers, policymakers, and quality assurance bodies. Machine learning can assist in the automated grading of quizzes, short answers, coding assignments, and writing assignments. However, human judgment will still be necessary for the assessment of creative and higher-order thinking skills. Validations of the automated grading systems should use rubrics and teacher grading. Additionally, developers should create systems that provide justifications for the grades given to students. Research studies should investigate the effectiveness of these systems. Finally, educational policymakers and quality assurance agencies should require all automated grading systems to be transparent, documented, accessible, private, non-biased, and subject to continual monitoring. Overall, the automated grading of educational assignments will remain a supported and supervised technology that enhances the efficiency of educators while maintaining the validity, accountability, fairness, and responsibility that teachers have in the educational process.

Limitations and Future Research

Despite the strengths of the review, there are some limitations to the study that should be considered. For instance, the study employed a qualitative systematic synthesis rather than a statistical meta-analysis to review the studies, as the studies that were included in the review had all subtly-differing designs and methods. Furthermore, the study only included English-language, peer-reviewed journal articles that were published between 2019 and 2025. Additionally, the quality of the research studies that were reviewed was not uniform; some studies published within the review included more information regarding datasets, validation, and study components than others. Finally, the review indicated that automated grading systems are best suited to the types of tasks that are typically more structured (such as short answer and programming questions) as opposed to tasks that require more creativity, critical thinking, or complex skills. Future research in this area can employ longitudinal studies that include multiple institutions and different cultures, include comparisons to graders who assess the submissions of students, and utilize different groups of learners. Additionally, future research can utilize the suggestions regarding specific aspects of automated grading systems that should be evaluated.

Conclusion

This review concludes that machine learning-enabled automated assessment and grading can become a valuable asset for educational practice if implemented with care. The evidence from the reviewed studies indicates that automated grading systems are most effective in structured and semi-structured assessment situations, such as short-answer questions, programming tasks, and writing assessments that use standardized rubrics. They offer significant workload reduction, rapid feedback, and scalable assessment opportunities. They should not, however, be considered independent arbiters of an assessment's value because the assessment of complex writing, creativity, critical thinking, originality, reasoning quality, programming style, and diverse learners requires expertise. The review findings also demonstrate that accuracy is not the only consideration for responsible deployment. Automated assessments must also be accompanied by validation, human-machine agreement, reliability, fairness, explainability, transparency, and contextual validation. Therefore, educational institutions should deploy automated assessment in a decision-support system format under human supervision, with teachers remaining responsible for moderation, feedback interpretation, appeals, and grading decisions. The review also identifies key areas for further empirical investigation: fairness, quality of feedback, learner trust, subgroup performance, classroom implementation, and learning outcomes over time. Automated assessment has excellent promise; however, its benefits lie in careful design and local validation, ethical governance, and ongoing human oversight.

References

[1] Alqurashi, A., Alharbi, B., & Sabbeh, S. (2025). An automated grading system for Arabic language Short-Answer questions using Deep learning. Engineering Technology & Applied Science Research, 15(5), 26665–26675. https://doi.org/10.48084/etasr.10917

[2] Andersen, N., Mang, J., Goldhammer, F., & Zehner, F. (2025). Algorithmic fairness in automatic short answer scoring. International Journal of Artificial Intelligence in Education, 35(5), 3128–3165. https://doi.org/10.1007/s40593-025-00495-5

[3] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[4] Beseiso, M., & Alzahrani, S. (2020). An empirical analysis of BERT embedding for automated essay scoring. International Journal of Advanced Computer Science and Applications, 11(10). https://doi.org/10.14569/ijacsa.2020.0111027

[5] Beseiso, M., Alzubi, O. A., & Rashaideh, H. (2021). A novel automated essay scoring approach for reliable higher educational assessments. Journal of Computing in Higher Education, 33(3), 727–746. https://doi.org/10.1007/s12528-021-09283-1

[6] Cisneros-González, J., Gordo-Herrera, N., Barcia-Santos, I., & Sánchez-Soriano, J. (2025). JorGPT: Instructor-Aided Grading of Programming Assignments with Large Language Models (LLMs). Future Internet, 17(6), 265. https://doi.org/10.3390/fi17060265

[7] Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

[8] Combéfis, S. (2022). Automated Code Assessment for Education: review, classification and perspectives on techniques and tools. Software, 1(1), 3–30. https://doi.org/10.3390/software1010002

[9] Conijn, R., Kahr, P., & Snijders, C. (2023). The effects of explanations in automated essay scoring systems on student trust and motivation. Journal of Learning Analytics, 10(1), 37–53. https://doi.org/10.18608/jla.2023.7801

[10] Deepshikha, D. (2025). A comprehensive review of AI-powered grading and tailored feedback in universities. Discover Artificial Intelligence, 5(1). https://doi.org/10.1007/s44163-025-00517-0

[11] Delgado-Pérez, P., & Medina-Bulo, I. (2020). Customizable and scalable automated assessment of C/C++ programming assignments. Computer Applications in Engineering Education, 28(6), 1449–1466. https://doi.org/10.1002/cae.22317

[12] Faseeh, M., Jaleel, A., Iqbal, N., Ghani, A., Abdusalomov, A., Mehmood, A., & Cho, Y. (2024). Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy. Mathematics, 12(21), 3416. https://doi.org/10.3390/math12213416

[13] Hannah, L., Jang, E. E., Shah, M., & Gupta, V. (2023). Validity Arguments for automated essay scoring of young students' writing traits. Language Assessment Quarterly, 20(4–5), 399–420. https://doi.org/10.1080/15434303.2023.2288253

[14] Hong, Q. N., Gonzalez-Reyes, A., & Pluye, P. (2018). Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). Journal of Evaluation in Clinical Practice, 24(3), 459–467. https://doi.org/10.1111/jep.12884

[15] Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: a literature review. PeerJ Computer Science, 5, e208. https://doi.org/10.7717/peerj-cs.208

[16] Kumar, V. S., & Boulanger, D. (2020). Automated essay scoring and the deep learning black box: How are rubric scores determined? International Journal of Artificial Intelligence in Education, 31(3), 538–584. https://doi.org/10.1007/s40593-020-00211-5

[17] Lohmann, J. F., Junge, F., Möller, J., Fleckenstein, J., Trüb, R., Keller, S., Jansen, T., & Horbach, A. (2025). Neural networks or linguistic features? - Comparing different Machine-Learning approaches for automated assessment of text quality traits among L1- and L2-Learners' argumentative essays. International Journal of Artificial Intelligence in Education, 35(3), 1178–1217. https://doi.org/10.1007/s40593-024-00426-w

[18] Mahdaoui, M., Nouh, S., Alaoui, M. S. E. K., & Kandali, K. (2025). Automated grading method of Python code submissions using large language models and machine learning. Information, 16(8), 674. https://doi.org/10.3390/info16080674

[19] Maslim, M., Wang, H. C., Putra, C. D., & Prabowo, Y. D. (2024). A trustworthy automated Short-Answer scoring system using a new dataset and hybrid transfer learning method. International Journal of Interactive Multimedia and Artificial Intelligence, 8(5), 37–45. https://doi.org/10.9781/ijimai.2024.02.003

[20] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

[21] Mekterović, I., Brkic, L., Milasinovic, B., & Baranovic, M. (2020). Building a comprehensive automated programming assessment system. IEEE Access, 8, 81154–81172. https://doi.org/10.1109/access.2020.2990980

[22] Messer, M., Brown, N. C. C., Kölling, M., & Shi, M. (2024). Automated Grading and Feedback Tools for Programming Education: A Systematic review. ACM Transactions on Computing Education, 24(1), 1–43. https://doi.org/10.1145/3636515

[23] Misgna, H., On, B., Lee, I., & Choi, G. S. (2024). A survey on deep learning-based automated essay scoring and feedback generation. Artificial Intelligence Review, 58(2). https://doi.org/10.1007/s10462-024-11017-5

[24] Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

[25] Ouahrani, L., & Bennouar, D. (2024). Paraphrase generation and supervised learning for improved automatic short answer grading. International Journal of Artificial Intelligence in Education, 34(4), 1627–1670. https://doi.org/10.1007/s40593-023-00391-w

[26] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

[27] Ramesh, D., & Sanampudi, S. K. (2021). An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review, 55(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

[28] Rojec, Ž., Puhan, J., & Fajfar, I. (2025). Practical Programming Exams with Automated Assessment Improve Student Engagement and Learning Outcomes. Education Sciences, 15(11), 1486. https://doi.org/10.3390/educsci15111486

[29] Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S., & Köller, O. (2019). Automated essay scoring at scale: a case study in Switzerland and Germany. ETS Research Report Series, 2019(1), 1–23. https://doi.org/10.1002/ets2.12249

[30] Uto, M. (2021). A review of deep-neural automated essay scoring models. Behaviormetrika, 48(2), 459–484. https://doi.org/10.1007/s41237-021-00142-y

[31] Xue, J., Tang, X., & Zheng, L. (2021). A hierarchical BERT-Based transfer Learning Approach for Multi-Dimensional Essay Scoring. IEEE Access, 9, 125403–125415. https://doi.org/10.1109/access.2021.3110683

[32] Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. Studies in Science Education, 56(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757

[33] Zhang, L., Huang, Y., Yang, X., Yu, S., & Zhuang, F. (2019). An automatic short-answer grading model for semi-open-ended questions. Interactive Learning Environments, 30(1), 177–190. https://doi.org/10.1080/10494820.2019.1648300

Conflict of Interest

The Authors declare that they have no known conflict of interest.

Funding

Authors declare no funding involved

Institutional Review Board Statement

Not applicable.