ISSN 2320 - 2602
Volume 15 No. 3, March 2026
International Journal of Advances in Computer Science and Technology
Available Online at http://www.warse.org/IJACST/static/pdf/file/ijacst011532026.pdf
https://doi.org/10.30534/ijacst/2026/011532026

Systematizing Research on AI-Based Automated Assessment: A Two-Axis Review Framework for Open-Ended Responses

Anastasia Vangelova[1]
[1]Technical University of Sofia, Bulgaria, avangelova@yahoo.com

Received Date : January 12, 2026 Accepted Date : February 10 , 2026 Published Date : March 07, 2026

ABSTRACT

Automated assessment of open-ended responses has evolved from rule-based and statistical approaches to large language model (LLM)-based and hybrid architectures. Despite this progress, the literature remains fragmented across scoring, formative feedback, contextual grounding, reliability, and system integration. This review proposes a two-axis classification framework that organizes existing research by research focus and level of implementation, from offline benchmark studies to learning management system (LMS)-integrated systems. The framework identifies six analytical categories and reveals an uneven pattern of development: scoring-oriented studies are methodologically more mature, while grounding, formative feedback, and institutional deployment often remain at the pilot stage. The review also highlights key gaps, including the lack of end-to-end architectures, limited reactive LMS integration, weakly structured outputs, and underdeveloped workflow-oriented perspectives. The central limitation of current research is not the absence of promising components, but their insufficient integration into coherent and reproducible assessment systems.

Key words : automated assessment, large language models, learning management systems, Retrieval-Augmented Generation.

1. INTRODUCTION

Automated assessment of open-ended student responses lies at the intersection of educational science and artificial intelligence. Unlike selected-response tests, where evaluation can rely on predefined correct answers, open-ended tasks require interpretation of argumentation, logical coherence, conceptual depth, and cognitive complexity [1], [2]. This interpretive nature makes such tasks pedagogically valuable, but also difficult to assess in a scalable, consistent, and transparent way.

Recent developments in large language models (LLMs) have significantly expanded the possibilities of automated assessment. Compared with earlier approaches based on predefined features or sequential neural representations, LLMs support more complex semantic and logical interpretation of student responses [3], [4]. At the same time, their use raises persistent concerns related to hallucinations, output instability, prompt sensitivity, limited interpretability, and insufficient grounding in instructional context.

Despite this rapid development, the literature on AI-based automated assessment remains highly fragmented. Existing studies range from statistical analyses of agreement between model outputs and human ratings to more complex systems integrated into institutional learning management system (LMS) environments. Although many of these studies rely on similar technological foundations, they differ substantially in pedagogical focus, validation methodology, and degree of implementation, which makes direct comparison difficult [5], [6].

In response, the present review proposes a two-axis classification framework that organizes the literature according to two dimensions: the primary research focus and the level of implementation. The first axis covers directions such as scoring, formative feedback, contextual grounding, reliability, and LMS integration, while the second distinguishes between offline benchmark studies, pilot deployments, and fully integrated LMS-based systems. The purpose of this framework is to clarify the current structure of the field and reveal the research gaps that continue to limit the emergence of robust and institutionally deployable AI-based assessment systems for open-ended responses.

2. BACKGROUND: EVOLUTION OF AUTOMATED ASSESSMENT

The development of automated assessment can be understood as a transition from explicitly controlled but semantically limited approaches toward more flexible and context-sensitive models.

The first generation emerged with systems such as Project Essay Grade (PEG), later followed by e-rater®, IntelliMetric™, and Intelligent Essay Assessor (IEA) [7]. These systems relied on explicitly formulated rules and manually selected textual features. Their strengths were transparency and reproducibility, but they were difficult to adapt to free and conceptually complex responses. Dependence on handcrafted rules and ontologies limited transfer across disciplines, languages, and task types, while surface features often had disproportionate influence on scores [8].

The second generation shifted toward statistical and feature-based prediction models trained on extracted linguistic indicators [4], [7], [9]. Student responses were represented through lexical, syntactic, discourse, and semantic features, and scores were modeled using regression and related statistical methods [4]. These models were scalable and relatively explainable, but remained dependent on indirect proxies of quality, vulnerable to gaming strategies, and difficult to generalize across topics, genres, and languages [2], [9].

The third generation introduced a more explicit machine learning formulation of automated assessment as a supervised task trained on expert-scored corpora [2], [9]. The focus moved toward nonlinear predictive algorithms such as Support Vector Machines, Random Forest, and Gradient Boosting [4]. These models often achieved high agreement with human scoring and strong scalability, but performance still depended heavily on manual feature design, partial semantic understanding, and retraining for new contexts [3].

The fourth generation introduced deep learning, shifting automated assessment away from handcrafted features toward learned internal representations derived directly from raw text [4], [10]. CNN, LSTM, BiLSTM, GRU, and attention-based architectures improved the modeling of local patterns, sequential structure, and contextual dependencies in student responses [2], [9]. These models reduced dependence on manual feature engineering and often improved benchmark performance, but introduced new limitations related to interpretability, dependence on large annotated datasets, and higher computational requirements [9], [11].

Across these generations, a consistent pattern emerges: early systems offered stronger formal control but limited semantic flexibility, while later approaches improved scalability and representational power at the cost of transparency, portability, or data efficiency. These limitations prepared the ground for the fifth generation, dominated by transformer-based architectures and large language models.

3. CONTEMPORARY LLM-BASED ASSESSMENT AND ITS ARCHITECTURAL EXTENSIONS

The fifth generation of automated assessment is dominated by transformer-based architectures and LLMs, which enable more complex semantic and logical interpretation of student responses through self-attention [3], [4]. In this setting, the model no longer functions only as a tool for similarity measurement or score prediction, but increasingly as an evaluator that interprets tasks, relates responses to criteria, and produces argued judgments.

Within automated assessment, two main functional roles can be distinguished. Generative LLMs act directly as virtual evaluators that produce scores, explanations, and feedback from a task, rubric, and student response. Encoder transformers generate contextual embeddings and function as semantic measurement tools or as inputs to downstream models. This distinction is methodologically important because it reflects whether the model performs the evaluation itself or supports a broader assessment pipeline.

A central mechanism for stabilizing LLM-based assessment is prompt engineering. Model behavior depends not only on architecture, but also on how the task, role, criteria, and output format are specified [12]. The literature shows a progression from minimal zero-shot instructions to more structured prompting strategies that include examples, intermediate reasoning steps, and analytical rubrics [13], [14]. In this context, rubric-guided evaluation is particularly important because explicit criteria constrain interpretation, improve consistency, and support criterion-level outputs. Structured formats such as JSON further improve reproducibility and downstream integration [15]. However, prompt-based control has clear limits: even small changes in wording may affect scores, and without grounding, models may still generate plausible but inaccurate interpretations [5], [6], [16].

These limitations are especially visible when LLMs are used as standalone evaluators. The literature consistently identifies four major weaknesses: lack of contextual grounding, probabilistic instability, limited interpretability, and risk of bias or manipulation [6], [16]. Hallucinations may introduce concepts not present in the student response, undermining construct validity [12], [17]. Outputs may vary across runs or model updates, and even generated explanations may not faithfully reflect the actual basis of a score [5], [15], [17].

One major response to these weaknesses is Retrieval-Augmented Generation (RAG). RAG combines generative LLMs with information retrieval in order to reduce hallucinations and ground assessment in external instructional material [18]. Instead of relying only on internal model parameters, the system retrieves relevant passages through semantic search based on embeddings and vector similarity, then inserts them into the prompt before generation [18]. In this way, evaluation becomes tied to a more concrete and verifiable instructional context.

A further extension appears in tool-augmented architectures. In such approaches, the LLM functions as a cognitive core, while critical operations are stabilized through

external tools and formal control mechanisms [19]. Calculators, interpreters, and validators can reduce arithmetic and logical errors, making the workflow more modular and verifiable [19].

This leads directly to the role of the learning management system (LMS) as the institutional layer of automated assessment. LMS platforms define the pedagogical and procedural context in which automated evaluation becomes valid: course structure, assignment configuration, rubric, grade publication rules, and traceability. Unlike autonomous AI tools operating outside the instructional process, LMS-based systems embed evaluation within predefined educational criteria and institutional regulations. In this sense, contemporary LLM-based assessment is shaped not only by the language model itself, but by the architectural structures built around it, including prompts, rubrics, contextual grounding, external tools, and LMS integration.

4. NEED FOR A NEW CLASSIFICATION FRAMEWORK

Despite the rapid growth of research on automated assessment and AI-based feedback, the literature remains highly fragmented. Existing studies range from statistical analyses of agreement between model outputs and human raters to more complex systems integrated into learning management systems. Although many of these works rely on similar technological foundations, they differ substantially in pedagogical focus, validation methodology, and degree of implementation, which makes direct comparison difficult [8]. This fragmentation is not only a matter of thematic breadth. It also reflects the fact that current studies develop along partially overlapping lines: some focus on scoring accuracy, others on formative feedback, others on RAG-based grounding, reliability, or LMS integration. As a result, the field cannot be adequately understood through a single organizing principle such as model type or task type alone.

The problem becomes even clearer when the unresolved gaps are considered. Much of the literature still lacks end-to-end assessment designs, reactive LMS integration, stable contextual grounding, machine-readable outputs, and a workflow-oriented view of evaluation. Without a clear framework, technically similar studies may be grouped together even when they differ substantially in pedagogical purpose and implementation maturity.

For this reason, the present review adopts a two-axis classification framework. The first axis captures the primary research focus of a study, while the second captures its level of implementation, ranging from offline benchmark experiments to full LMS-based deployment. This makes it possible to distinguish more clearly between experimental scoring studies and practically deployable assessment systems.

5. PROPOSED TWO-AXIS CLASSIFICATION FRAMEWORK

To address the fragmentation of the literature, this review adopts a two-axis classification framework that organizes existing studies according to both their primary contribution and their degree of practical implementation. The framework captures a central structural feature of current research: studies built on similar technological foundations may still differ substantially in pedagogical purpose, validation logic, and institutional maturity.

The first dimension, Axis 1 (Research Focus), captures the dominant purpose of a given study. In the present framework, this axis includes five major directions: Scoring, Feedback, Grounding, Reliability, and LMS Integration. Scoring covers studies primarily concerned with the accuracy of AI-generated scores in relation to human ratings. Feedback includes work focused on formative or pedagogically supportive AI-generated responses. Grounding refers to research that restricts evaluation through mechanisms such as Retrieval-Augmented Generation. Reliability includes studies concerned with consistency, variability, agreement, and robustness of model behavior. LMS integration covers research that moves beyond isolated evaluation toward institutionally embedded systems operating within learning management platforms.

The second dimension, Axis 2 (Level of Implementation), captures the practical maturity of the research. This dimension distinguishes between offline benchmark evaluation, controlled pilot deployment, and full LMS integration in a real educational setting. This distinction is essential because a study may be methodologically strong in model performance while remaining far from real educational deployment, whereas another may contribute less to benchmark comparison but more to institutional applicability.

The combination of these two axes produces a matrix that structures the literature into six analytical categories (A–F). These categories are not intended as rigid classes, but as review categories that make dominant patterns in the literature more visible and comparable.

Category A includes offline experimental studies on public corpora, such as ASAP or TOEFL11, where the main objective is to measure agreement between AI-generated and human scores using metrics such as quadratic weighted kappa (QWK) or Pearson correlation [1], [9]. These studies often show strong statistical performance, but remain detached from real educational workflows and rarely address institutional integration or long-term stability [17].

Category B extends the analysis to authentic student data and domain-specific rubrics in controlled educational settings [20]. These studies improve ecological validity, but most remain at the prototype stage and do not yet provide automated

production workflows or long-term evidence of pedagogical impact [21].

Category C captures studies that introduce contextual grounding through RAG in order to reduce hallucinations and constrain evaluation to instructional materials [22], [23]. Although these approaches improve factual consistency through vector retrieval, they often remain pilot implementations without full LMS automation or detailed analysis of scalability [6], [19].

Category D includes studies centered on formative feedback, often linked to Bloom's taxonomy or dialogic pedagogical strategies [12], [24]. These studies frequently report high perceived usefulness, but robust evidence for long-term learning improvement and systematic institutional integration remains limited [8], [25], [26].

Category E covers reliability-oriented studies that investigate the consistency and variability of LLM outputs using expert judgment and statistical analysis [1], [27]. Such work is critical for identifying methodological risks, but it is rarely linked to a real educational workflow or LMS-driven assessment process.

Category F represents the most mature line of development: integrated intelligent systems using API-based communication and automated workflows inside LMS environments [11], [22], [28]. These solutions come closest to institutional deployment, yet even here important limitations remain, including lack of standardized interoperability, limited human-in-the-loop control, and insufficient analysis under high-load conditions [18].

Overall, the framework reveals a clear developmental trajectory in the literature: from isolated offline experiments toward increasingly integrated systems with institutional relevance. At the same time, it shows that a large part of the field remains concentrated in offline or pilot stages, without full automation or stable LMS integration. The overall conceptual structure of the framework is illustrated in Table 1, and the comparative synthesis of the six categories is presented in Table 2.

Table 1: Two-axis classification matrix (Axis 1 × Axis 2) of research on AI-based automated assessment

Axis 1 / Axis 2	Axis 1: Focus			
	Scoring	Feedback	Grounding (RAG)	LMS System Integration
Full LMS Integration				**F**
Controlled Pilot Deployment	**B**	**D**	**C**	
Offline Benchmark Evaluation	**A**			**E**

Table 2: Classification of research in automated assessment by Axis 1 and Axis 2

Category	Axis 1: Research Focus	Axis 2: Level of Implementation	Primary Objective	Typical Methods	Key Limitations
A	Scoring	Offline benchmark evaluation	Measuring the accuracy of AI-based assessment against human scoring	QWK, Pearson r, zero-/few-shot prompting, public corpora (ASAP, TOEFL11)	Lack of a real LMS environment; lack of pedagogical impact; lack of analysis of the long-term stability of models
B	Scoring	Controlled pilot deployment	Validating automated assessment on real educational data	Domain-specific corpora, rubrics, analysis of inter-rater variability	Lack of a usable production system; absence of real-time assessment; lack of evidence on long-term educational impact
C	Contextual grounding (RAG)	Pilot deployments	Contextually grounding assessment through instructional materials	RAG architectures, vector databases, top-k retrieval, human-in-the-loop	Lack of built-in LMS automation; limited scalability; use of outdated instructional content
D	Formative feedback	Pilot deployments	Improving learning through formative AI-generated feedback	Pedagogical prompts, Likert scales, Bloom's taxonomy, Socratic dialogue	Lack of long-term learning outcomes; lack of institutional integration
E	Reliability	Offline benchmark evaluation	Analyzing the reliability and stability of AI-generated content	Expert evaluations, ANOVA, t-test, subjective metrics	Lack of adaptive assessment; absence of integration into the LMS assessment workflow; need for manual interaction with the language model
F	System integration (LMS)	LMS integration	Building end-to-end intelligent assessment systems	Modular architectures, API integration, automated processes	Lack of human-in-the-loop control; lack of use of a standard, widely recognized protocol for LMS integration; limited analysis of system behavior under large-scale load

6. RESEARCH GAPS REVEALED BY THE FRAMEWORK

The proposed framework makes the main research gaps in the field more visible. Although the literature has clearly progressed from isolated offline evaluation toward more complex and institutionally relevant solutions, this progression remains incomplete. A large part of the field still occupies offline or pilot stages, while fully automated, institutionally embedded, and methodologically reproducible systems remain relatively rare.

The first major gap is the lack of end-to-end assessment architectures. In many studies, the model operates according to an "input-to-score" logic and is not embedded in the full lifecycle of an assessment event. Stages such as submission, contextual retrieval, structured result storage, formal recording, and traceability are often absent. This limits the practical value of otherwise technically strong studies, since real educational assessment unfolds as a multi-stage pedagogical and administrative process.

A second gap concerns reactive LMS integration. Even when platforms are included, they often function only as repositories or display interfaces. Event-driven architectures in which assessment is triggered automatically by instructional events remain underexplored. Polling-based approaches, manual intermediate steps, and weak orchestration logic increase latency and reduce the sustainability of deployment.

A third gap involves the stability of contextual grounding. RAG has become a central strategy for reducing hallucinations and aligning evaluation with instructional materials, but most grounding-oriented studies remain at the pilot stage. In many cases, the knowledge base is static, while dynamic updating of learning resources, retrieval quality control, and version management are not formalized as part of a production architecture.

A fourth gap concerns structured and reproducible outputs. Across multiple categories, results are frequently presented as single scores or free-text justifications without machine-readable formalization. This weakens automatic verification, comparison across runs or versions, and criterion-level analysis. Closely related to this is the continued concentration of the literature in English-language and high-resource contexts. Transfer to smaller languages, domain-specific settings, and bilingual educational environments remains insufficiently explored.

More broadly, the framework shows that automated assessment is still too often treated as a model-centered task rather than as a workflow-centered institutional process. Mechanisms for exception handling, repeated calls, state management, publication control, and behavior under high-load conditions are frequently weakly specified or absent. The central problem is therefore not the absence of promising components, but their insufficient integration into unified, traceable, and reproducible architectures.

7. CONCLUSION

This review examined the development of automated assessment of open-ended responses from early rule-based and statistical models to contemporary transformer-based large language models and hybrid architectures. The analysis showed that the current landscape is best understood as hybrid: generative LLMs expand the semantic and argumentative reach of automated assessment, while encoder-based models, contextual grounding, rubrics, external tools, and LMS-oriented workflows provide structure and control.

The central contribution of this article is the proposed two-axis classification framework, which organizes the literature according to both research focus and level of implementation. Through this framework, it becomes easier to distinguish between scoring-oriented, feedback-oriented, grounding-focused, reliability-oriented, and LMS-integrated studies, while also identifying how far each line of work has progressed toward real educational deployment. The review shows that the main limitation of current research is not the lack of promising approaches, but their insufficient integration into coherent, traceable, and reproducible assessment architectures. Sustainable automated assessment therefore depends on combining contemporary language models with relevant instructional context, analytical rubrics, algorithmic control, and stable embedding into real educational workflows.

ACKNOWLEDGEMENT

The author used ChatGPT as an assistive tool for language editing, grammar correction, and improving the clarity and coherence of the manuscript. All scientific content, interpretation, structure of the review, and final decisions remain the sole responsibility of the author.

REFERENCES

1. E. Emirtekin. **Large language model-powered automated assessment: A systematic review**, *Applied Sciences*, vol. 15, no. 10, p. 5683, 2025.
2. R. Gao, H. E. Merzdorf, S. Anwar, M. C. Hipwell, and A. R. Srinivasa. **Automatic assessment of text-based responses in post-secondary education: A systematic review**, *Computers and Education: Artificial Intelligence*, vol. 6, p. 100206, 2024.
3. Q. Wang. **A multifaceted architecture to automate essay scoring for assessing English article writing: integrating semantic, thematic, and linguistic representations**, *Computers and Electrical Engineering*, vol. 118, p. 109308, 2024.
4. X. Tang, H. Chen, D. Lin, and K. Li. **Harnessing LLMs for multi-dimensional writing assessment: reliability**

and alignment with human judgments, *Heliyon*, vol. 10, no. 14, p. e34262, 2024.
5. M. D. Shermis. **Using ChatGPT to score essays and short-form constructed responses**, *Assessing Writing*, vol. 66, p. 100988, 2025.
6. P. C. Mendonça, F. Quintal, and F. Mendonça. **Evaluating LLMs for automated scoring in formative assessments**, *Applied Sciences*, vol. 15, no. 5, p. 2787, 2025.
7. S. Dikli. **An overview of automated scoring of essays**, *Journal of Technology, Learning, and Assessment*, vol. 5, no. 1, 2006. [Online]. Available: http://www.jtla.org. Accessed on: Dec. 18, 2025.
8. E. Del Gobbo, A. Guarino, B. Cafarelli, L. Grilli, and P. Limone. **Automatic evaluation of open-ended questions for online learning: A systematic mapping**, *Studies in Educational Evaluation*, vol. 77, p. 101258, 2023.
9. J. Sun, T. Song, W. Peng, and J. Song. **A survey of automated essay scoring: Challenges, advances, and future**, *Neurocomputing*, vol. 650, p. 130916, 2025.
10. Y. Wang, J. Huang, L. Du, Y. Guo, Y. Liu, and R. Wang. **Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity**, *Computers and Education: Artificial Intelligence*, vol. 9, p. 100481, 2025.
11. S. Mahamad, Y. H. Chin, N. I. N. Zulmuksah, M. M. Haque, M. Shaheen, and K. Nisar. **Technical review: Architecting an AI-driven decision support system for enhanced online learning and assessment**, *Future Internet*, vol. 17, no. 9, p. 383, 2025.
12. L. J. Jacobsen and K. E. Weber. **The promises and pitfalls of large language models as feedback providers: A study of prompt engineering and the quality of AI-driven feedback**, *AI*, vol. 6, no. 2, p. 35, 2025.
13. G.-G. Lee, E. Latif, X. Wu, N. Liu, and X. Zhai. **Applying large language models and chain-of-thought for automatic scoring**, *Computers and Education: Artificial Intelligence*, vol. 6, p. 100213, 2024.
14. T. A. F. E. Nkoyo, C. F. Ijezue, A. I. Amjad, M. Amjad, S. Butt, and G. Castañeda-Garza. **Advances in auto-grading with large language models: A cross-disciplinary survey**, in *Proc. 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pp. 477–498.
15. F. García-Varela, M. Nussbaum, M. Mendoza, C. Martínez-Troncoso, and Z. Bekerman. **ChatGPT as a stable and fair tool for automated essay scoring**, *Education Sciences*, vol. 15, no. 8, p. 946, 2025.
16. J. S. Jauhiainen and A. G. Guerra. **Evaluating students' open-ended written responses with LLMs: Using the RAG framework for GPT-3.5, GPT-4, Claude-3, and Mistral-Large**, *Advances in Artificial Intelligence and Machine Learning*, vol. 4, no. 4, pp. 3097–3113, 2024.
17. A. Pack, A. Barrett, and J. Escalante. **Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability**, *Computers and Education: Artificial Intelligence*, vol. 6, p. 100234, 2024.
18. I. Papachristou, G. Dimitroulakos, and C. Vassilakis. **Automated test generation and marking using LLMs**, *Electronics*, vol. 14, no. 14, p. 2835, 2025.
19. W. Villegas-Ch, R. Gutierrez, J. García-Ortiz, and V. Guevara. **Explainable educational assistant integrated in Moodle: Automated semantic assessment and adaptive tutoring based on NLP and XAI**, *Discover Artificial Intelligence*, vol. 5, no. 1, p. 191, 2025.
20. J. Y. Jung, L. Tyack, and M. Von Davier. **Towards the implementation of automated scoring in international large-scale assessments: Scalability and quality control**, *Computers and Education: Artificial Intelligence*, vol. 8, p. 100375, 2025.
21. J. P. Bernius, S. Krusche, and B. Bruegge. **Machine learning based feedback on textual student answers in large courses**, *Computers and Education: Artificial Intelligence*, vol. 3, p. 100081, 2022.
22. D.-M. Córdova-Esparza. **AI-powered educational agents: Opportunities, innovations, and ethical challenges**, *Information*, vol. 16, no. 6, p. 469, 2025.
23. L. Krupp, J. Bley, I. Gobbi, A. Geng, S. Müller, *et al.* **LLM-generated tips rival expert-created tips in helping students answer quantum-computing questions**, *EPJ Quantum Technology*, vol. 12, no. 1, p. 33, 2025.
24. Y. Chen, Y. Li, Y. Ren, Y. Liu, and Y. Ma. **Educational evaluation with MLLMs: Framework, dataset, and comprehensive assessment**, *Electronics*, vol. 14, no. 18, p. 3713, 2025.
25. A. V. Y. Lee. **Supporting students' generation of feedback in large-scale online course with artificial intelligence-enabled evaluation**, *Studies in Educational Evaluation*, vol. 77, p. 101250, 2023.
26. J. Wilson, C. Ahrendt, E. A. Fudge, A. Raiche, G. Beard, and C. MacArthur. **Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation**, *Computers & Education*, vol. 168, p. 104208, 2021.
27. V. Hackl, A. E. Müller, M. Granitzer, and M. Sailer. **Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings**, in *Frontiers in Education*, vol. 8, p. 1272229, Dec. 2023.
28. S. Wang, F. Wang, Z. Zhu, J. Wang, T. Tran, and Z. Du. **Artificial intelligence in education: A systematic literature review**, *Expert Systems with Applications*, vol. 252, p. 124167, 2024.