Computer Science Review 61 (2026) 100929

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

How does leveraging artificial intelligence in assessments impact student outcomes? a systematic review

Mohammed Bashraheel [a,b,*], Gheorghita Ghinea [a]

[a] Computer Science Department, Brunel University London, Kingston Lane, Uxbridge, UB8 3PH, MX, United Kingdom
[b] Department of Computer Science, College of Engineering and Computer Science, Jazan University, Al Maarefah Rd, Jazan, 45142, Jazan, Saudi Arabia

HIGHLIGHTS

- Review of 159 studies on AI in educational assessments (1997–2024).
- Classification is the most used AI technique for student performance prediction.
- Formative assessments dominate over summative assessments in AI-supported studies.
- AI tools enhance feedback, engagement, and learning outcomes in higher education.
- A thematic overview of AI assessment types and outcomes is provided to guide future research.

ARTICLE INFO

Keywords:
Machine learning
Deep learning
Large language models
Formative assessment
Summative assessment
Higher education

ABSTRACT

Advancements in Artificial Intelligence (AI) are having a profound impact across numerous domains, including education, particularly in the area of assessment. Within higher education, AI-based assessment has gained increasing attention for its potential to enhance student learning processes and outcomes. Following PRISMA guidelines and covering research published between 1997 and 2024, this systematic literature review (SLR) analyzes 159 studies that apply AI techniques, including machine learning (ML), deep learning (DL), and large language models (LLMs), in formative and summative assessment contexts to predict student outcomes. The findings indicate that, while AI integration can enhance assessment strategies and learning outcomes, classification-based models dominate the literature, and more than 80% of studies rely on private or institution-specific datasets, limiting reproducibility and large-scale validation. This review offers a comprehensive comparative synthesis of AI-driven formative and summative assessment approaches in higher education, highlighting methodological trends, evidence, and research gaps.

1. Introduction

In recent decades, the development of educational assessment methods has drawn increasing attention to the need for synthesizing and disseminating global research in the field [26]. The evolution of assessment principles, policies, and practices has marked significant milestones in higher education, creating an opportune moment to reflect on existing contributions and identify areas requiring further investigation. Despite these advancements, a persistent gap remains in research examining how assessment techniques and feedback practices are designed, implemented, and interpreted within modern educational environments, particularly in higher education [51]. This gap limits the development of effective assessment policies and restricts broader insights into how assessment practices shape instructional strategies and student learning outcomes [26].

Educational assessment generally involves the systematic collection of evidence to evaluate student learning in relation to predefined learning objectives [16,56]. Such evidence may be derived from assessments embedded in coursework, class activities, or standardized evaluations and is used to diagnose students' knowledge, skills, and learning progress. Assessment practices also support instructional decision-making by identifying strengths, weaknesses, and areas requiring intervention [28,51]. While assessment focuses on evidence collection, evaluation involves interpreting and judging the quality of student achievement based on that evidence [17,92]. In higher education, these

[*] Corresponding author at: Computer Science Department, Brunel University London, Kingston Lane, Uxbridge, UB8 3PH, MX, United Kingdom.
Email addresses: mohammedsalema.bashraheel@brunel.ac.uk, msbashraheel@jazanu.edu.sa (M. Bashraheel), george.ghinea@brunel.ac.uk (G. Ghinea).

https://doi.org/10.1016/j.cosrev.2026.100929
Received 12 June 2025; Received in revised form 2 February 2026; Accepted 6 February 2026
Available online 12 February 2026

Table 1
Comparison of formative and summative assessment characteristics.

Characteristics	Formative assessment	Summative assessment
Role	Assessment for learning	Evaluation of learning
Impact	Low stakes	High stakes
Aim	For learning and curriculum enhancement	For final judgment
Measurement	Based on the learning process	Based on overall performance
Stage	At the beginning, during, or after a task	At the end of a task
Grade Status	No	Yes
Feedback	Ongoing to identify strengths and weakness	Overall progress
Approach	Qualitative	Quantitative
Example	General questions during a class	Final exam

processes are closely interrelated and collectively inform academic decision-making and accountability practices [17].

Assessment practices are commonly categorized into formative and summative types based on their purposes and timings [57,142]. Formative assessment supports learning during the instructional process by providing timely feedback-oriented information that can be used to adapt teaching and guide student improvement [24]. Typically low-stakes in nature, formative assessment emphasizes diagnosis and continuous feedback through activities such as drafts, in-class tasks, and peer or instructor feedback. Its role in curriculum design and instructional improvement is widely acknowledged [30,143]. Summative assessment, by contrast, evaluates student learning at the conclusion of a learning period and is often high-stakes, contributing substantially to final grades and institutional evaluation [37,62]. Common examples include final examinations, projects, and cumulative tests. Although recent perspectives emphasize the formative value embedded within all assessment practices [130], stating that “the current accepted theory no longer separates formative-summative assessment and, what is more, requires all assessments to be primarily formative in nature,” summative assessment continues to serve a critical role in certifying overall achievement relative to benchmark learning objectives in higher education.

Formative and summative assessments differ primarily in their objectives, timing, and impact on learning [25,43,131,146]. While formative assessment prioritizes feedback and instructional adjustment, summative assessment provides an overall measure of competency at the end of a teaching period. Both play complementary roles in higher education, yet a single assessment approach may not adequately address both purposes simultaneously. Table 1 summarizes these key distinctions and highlights the importance of aligning assessment methods with their intended educational goals.

Recent advances in digital technologies and computing have facilitated the growing adoption of intelligent systems in higher education. In particular, advances in AI, including ML, DL, and LLMs, as well as educational data mining, have expanded the possibilities for the automation of assessment, feedback generation, and learning analytics [8]. These technologies offer new opportunities to enhance assessment quality, scalability, and personalization within higher education contexts.

A growing body of SLRs has examined assessment practices in higher education, AI applications in education, and AI-driven assessment from different perspectives. Some reviews concentrate on formative assessment and feedback in higher education without explicitly considering the role of AI [95], while others investigate AI applications in higher education more broadly, with assessment and evaluation addressed as one of several application areas [148]. More recent SLRs [54,120,149,150] have focused on AI-based or automated assessment systems; however, these studies often emphasize particular assessment formats or disciplinary contexts, such as text-based assessment in post-secondary education or science and STEM assessment, and typically consider individual AI paradigms, such as ML [149] or NLP [54] in isolation. In addition, emerging reviews on educational technologies and LLMs often adopt a broad emerging technology perspective without a systematic comparison of AI techniques in assessment [120], or focus on restricted educational contexts such as K-12 education rather than higher education [150].

As summarized in Table 2, existing SLRs collectively offer an uneven view of AI-driven assessment research. In particular, relatively few reviews explicitly distinguish between formative and summative assessment purposes as analytical dimensions, and limited attention is given to systematically comparing multiple AI paradigms, including ML, DL, and LLMs, within a single assessment-focused framework. Moreover, evidence regarding reported effectiveness and data sources is often synthesized inconsistently across higher education contexts. This limits the extent to which current reviews support a comprehensive understanding of how different AI techniques are applied to assessment purposes and how they relate to student learning outcomes in higher education. To address these limitations, this SLR synthesizes and analyzes quantitative and qualitative studies that apply ML, DL, and LLM techniques to both formative and summative assessments in higher education. By systematically comparing assessment purposes, AI techniques, reported effectiveness, and data sources, this review provides a structured synthesis of current research and identifies methodological, technical, and practical challenges to inform future research and practice.

The remainder of this paper is organized as follows. Section 2 describes the methods and materials used to conduct the SLR. Section 3 presents the results. Section 4 discusses the findings, implications, and future research directions, and Section 5 concludes with limitations and conclusions.

2. Methods and materials

This study employs an SLR methodology based on the PRISMA 2020 Statement [101,102] to identify and analyze recent formative and summative assessment and feedback in higher education using AI, including ML, DL, and LLM techniques. The review follows the three standard phases of an SLR: planning, conducting, and reporting.

This SLR aims to identify and synthesize peer-reviewed studies that apply formative and summative assessments and feedback mechanisms to enhance student learning outcomes. Specific research questions guide the review, and relevant data are extracted accordingly. To ensure coverage and relevance, all included publications were retrieved from Scopus, Web of Science, IEEE Xplore, and the ACM Digital Library, using the keywords described in Section 2.2.3.1.

Overall, the review methodology adheres to established approaches for analyzing how AI-driven techniques support learning processes, improve educational quality, and predict student outcomes. This SLR contributes by providing synthesized evidence, methodological insights, and future research directions on how AI techniques, powered by ML, DL, and LLMs, are applied to formative and summative assessments to support student learning and outcome prediction in higher education.

2.1. Planning the SLR

In line with PRISMA 2020 guidelines, the planning phase of this SLR focused on establishing a transparent and reproducible review protocol. Prior to conducting the review, existing SLRs were examined to assess overlap, refine the scope, and ensure that the planned review addressed limitations in the current evidence base.

Table 2 summarizes the comparison of relevant existing reviews and informs the definition of the review scope, particularly with respect to assessment purposes, formative and summative, AI paradigms, including ML, DL, and LLMs, and higher education contexts. This comparative analysis supports the formulation of focused research questions and analytical dimensions rather than serving as a restatement of the research motivation.

Table 2

Summary of prior review studies in relation to AI-driven assessment in higher education.

Study	Aim	Timeframe	Primary studies	AI scope	Focus
[54]	To review text-based automated assessment in post-secondary education.	2017–2023	93	NLP-based	Focuses on automated grading and feedback without comparative formative and summative analysis.
[150]	To review the use of ChatGPT in K–12 education.	2022–2023	13	ChatGPT (LLMs)	Emphasizes pedagogy, ethics, and stakeholder perspectives rather than assessment comparisons.
[149]	To review ML in science assessment.	2008–2018	49	ML	Emphasizes supervised learning and validity in science education.
[120]	To review emerging technologies for assessment and feedback in higher education	2016–2021	38	AI, Learning Analytics, XR	Focuses on applications and practices rather than AI paradigms or assessment purposes.
[95]	To review formative assessment and feedback in higher education	2000–2019	28	None	Pedagogical and causal evidence focus; no AI-based assessment.
[148]	To review AI applications in higher education	2007–2018	146	AI	Broad mapping of AI applications, not assessment-centric.
This Review	To review AI-driven formative and summative assessment in higher education	1997–2024	159	AI (ML, DL, LLMs)	Comparative, assessment-centric synthesis of AI-driven formative and summative assessment, analyzing AI techniques, effectiveness evidence, and data sources.

Table 3

PICO terms and definitions applied in this review.

PICO term	Summary
P: Population	Studies involving students in higher education settings where formative assessment, summative assessment, or both are used with or without feedback.
I: Intervention or Interest	Use of AI techniques, including ML, DL, and/or LLM-based approaches, to support formative and/or summative assessment processes (e.g., automated scoring, feedback generation, prediction of learning outcomes).
C: Comparison	Comparisons across assessment purpose, including formative vs summative vs both, and/or across AI techniques, such as ML vs DL vs LLM, as analyzed in this review.
O: Outcome	Reported student outcomes (e.g., performance and achievement, engagement, behavior) and assessment-related outcomes (e.g., prediction accuracy and evidence related to feedback quality and validity and reliability), including contextual factors and features influencing effectiveness.

Based on this planning process, a structured SLR protocol was developed, encompassing research question formulation, inclusion and exclusion criteria, comprehensive search strategy design, data extraction and synthesis procedures, quality assessment, and transparent reporting of findings.

2.2. *Conducting the SLR*

2.2.1. *Preparing the protocol*

Following the planning phase, the SLR protocol was operationalized to guide the execution of the review. The protocol specified the finalized research questions, inclusion and exclusion criteria, search strategy, study selection procedures, data extraction process, quality appraisal methods, and synthesis approach.

2.2.2. *Research questions*

As shown in Table 3, we used the PICO framework (Population, Intervention/Interest, Comparison, Outcomes) [74] to structure and refine the review questions and eligibility criteria.

Research in [125] suggests that clearly defined and focused research questions improve the efficiency and rigor of an SLR. Accordingly, this study is guided by four research questions, summarized in Table 4.

Table 4

Research questions and objectives.

Questions	Objectives
What is the current state of research on the implications of AI-driven formative and summative assessments and feedback for student outcomes in higher education?	To review the current state of research on the implications of AI-driven formative and summative assessments and feedback for student outcomes in higher education.
What AI techniques, including ML, DL, and LLM, are used for formative and summative assessments and feedback to enhance student outcomes?	To investigate AI techniques, including ML, DL, and LLM, used for formative and summative assessments and feedback to enhance student outcomes.
Is there any existing evidence regarding the effectiveness of AI-driven formative and summative assessment applications in predicting student outcomes?	To investigate the existing evidence regarding the effectiveness of AI-driven formative and summative assessment applications in predicting student outcomes.
What data sources are utilized for AI-driven assessments in higher education?	To discover what data sources are utilized for AI-driven assessments in higher education.

2.2.3. *Information sources and search strategy*

A systematic search was conducted on 29 July 2024 using four multidisciplinary databases: Scopus, Clarivate Web of Science, IEEE Xplore, and the ACM Digital Library. Scopus and Web of Science were selected for their comprehensive coverage of scholarly journals across disciplines [91,109], while IEEE Xplore and ACM Digital Library were chosen due to their prominence in computer science and engineering research [44,90,111].

2.2.3.1. Search term. To identify relevant studies, search terms were defined to capture concepts related to higher education, assessment, feedback, and AI technologies. Synonyms and alternative expressions were included to broaden coverage. The complete Boolean query is reported as follows: (universit* OR learning OR education*) AND ((formative OR assessment OR evaluat* OR exam* OR quiz* OR test* OR assignment OR homework) AND (feedback) AND (student)) AND (“active learning” OR “student engagement” OR “student outcome” OR “academic achievement” OR “learning outcome” OR “educational outcome” OR “learning objective” OR “learning progress*” OR “academic performance”) AND (chatbot* OR “ChatGPT” OR “Artificial intelligence” OR “Machine learning” OR “Deep learning” OR “large language model” OR AI OR “natural language processing”).

Table 5
Inclusion and exclusion criteria for the SLR.

Inclusion criteria	Exclusion criteria
Research articles applying AI techniques, including ML, DL, or LLM-based, to support formative assessment, summative assessment, or assessment-related feedback.	Studies not incorporating AI techniques, including ML, DL, or LLM-based, for formative assessment, summative assessments, or assessment-related feedback.
Studies published from 1997 to 2024 cover the earliest available publications and recent advancements.	Studies published before 1997.
Articles published in the English language to support consistent screening and reproducible data extraction.	Survey and review papers are excluded as they do not present original research or methodological approaches.
Only articles available in a full-text format to ensure quality and reliability of findings and methodologies.	Excludes grey literature (e.g., reports, white papers) due to lack of scientific validation.
Peer-reviewed journal articles and conference papers.	Exclude learning analytics studies unless they are explicitly part of AI-based formative or summative assessment or assessment-related feedback.
Focus on higher education settings.	Articles focusing on primary, secondary, or high school education or non-academic sources.
Studies reporting evidence on effectiveness, performance, or impact on student learning outcomes (quantitative metrics and/or qualitative evaluation).	Duplicate publications based on the same dataset (retain the most comprehensive version).

This transparent and systematic search strategy enhances reproducibility and ensures comprehensive retrieval of relevant studies.

2.2.4. Eligibility criteria

This section consists of the scope of this SLR, as shown in Table 5.

The inclusion and exclusion criteria were designed to ensure methodological rigor, relevance, and reproducibility, and to minimize bias arising from non-peer-reviewed, non-empirical, or contextually misaligned studies. The review focused on research articles that apply formative, summative, or both assessment strategies and use ML, DL, or LLM-based techniques to support student learning outcomes and feedback in higher education contexts. These methods represent the most recent and widely adopted AI technologies in educational research [148], particularly in higher education contexts. Restricting the scope to AI-based approaches ensures that the review captures state of the art assessment methodologies relevant to current computational and data-driven educational practices.

The literature search was limited to studies published between 1997 and 2024 to capture the emergence and evolution of AI- and ML-based approaches in educational assessment. Studies published before this period rarely incorporated ML-based techniques in educational assessment contexts due to limited computational capacity and the absence of large-scale digital educational data [22,117]. Subsequent years reflect substantial growth in AI-supported assessment research. Limiting the review to this period enables coverage of both foundational and recent advances while maintaining relevance to modern assessment practices.

Only peer-reviewed journal articles and conference papers published in full-text format were included to ensure scientific quality, transparency, and replicability. Conference papers were included alongside journal articles because high-impact and innovative research in computer science is frequently disseminated through reputable conference venues. Included studies were required to report empirical results, supported by quantitative evaluation metrics and/or qualitative evidence to enable the assessment of methodological effectiveness.

To maintain contextual consistency, the review focused exclusively on higher education settings and included quantitative, qualitative, and mixed-methods studies. Studies conducted in primary or secondary education, non-academic contexts, or domains unrelated to education were excluded, as they involve different pedagogical structures and learner characteristics that fall outside the scope of this review. The English language restriction was applied to ensure consistent screening and reproducible data extraction and to reduce the misclassification risk introduced by translation. Learning analytics studies were excluded unless they were directly tied to assessment processes to avoid conflating monitoring and prediction research with AI-driven assessment and feedback.

Survey and review papers were excluded to avoid duplication of evidence and because they do not provide original empirical findings or methodological validation [96]. Grey literature was excluded due to the lack of standardized peer review and potential limitations in methodological transparency [58]. In addition, theoretical studies without practical implementation or performance evaluation were omitted, as empirical validation is essential for evaluating AI-based assessment methods. Finally, where multiple publications reported results from the same dataset, only the most comprehensive version was retained to avoid double-counting evidence and inflating study weight.

2.2.5. Study selection process

The choice of published studies was a multi-step procedure. First, primary studies were identified from the four databases, consolidated into a single Excel file, and duplicate studies. Second, the studies were screened, categorizing them into relevant, doubtful, and irrelevant groups. In other words, a screening status column was added in the Excel file, which contained 0,1, and 2 values, where 0 meant not included, 1 represented the included paper, and 2 indicated the doubtful study [103]. This involved reviewing titles, abstracts, and keywords to refine the selection as an initial screening. As a single researcher, the screening was conducted and then sent to another researcher to assess the selected studies for potential bias. Then we reviewed the doubtful papers and sought a second author's perspective to reach a consensus on inclusion. After that, the rejected studies were maintained with reasons for the rejection [74]. We then checked the full texts of the included studies, excluding those without full texts. In addition, to investigate the effectiveness of the eligibility criteria, a random sample of the included papers was applied [107]. Moreover, all references were imported into the reference management software RefWorks for citation. Finally, we reapplied the eligibility criteria to all papers, excluded studies with documented reasons, and finalized the total set of primary studies, as detailed in Fig. 1. This multi-stage screening and verification process was designed to reduce selection bias, ensure consistency in study inclusion, and improve the reliability of the final sample.

2.2.6. Data collection process

2.2.6.1. Data extraction. A structured data extraction process was applied to enhance the rigor and reliability of the research synthesis, minimize bias, and ensure consistency in study selection. The following data were collected from the selected articles.

First, the author(s), year of publication, title, abstract, keywords, and publication information were extracted. These elements serve not only as formal citation information but also provide contextual background regarding the relevance and significance of each study [93].

Second, the types of assessments employed in each study, formative, summative, or both, were examined and documented, along with the AI-based feedback mechanisms used to support student learning outcomes. In addition, we systematically documented and categorized various forms of assessment, such as self-assessment and peer assessment.

Third, the educational domain associated with each assessment, such as computer science, engineering, and other fields, was recorded to contextualize experimental settings and applications. Furthermore, we recorded the types and sizes of data utilized in each study, as data characteristics play a critical role in the selection and performance of AI-based assessment methods. Feature sets used for predicting student outcomes were also recorded to identify commonly employed predictors and to inform recommendations for future research.

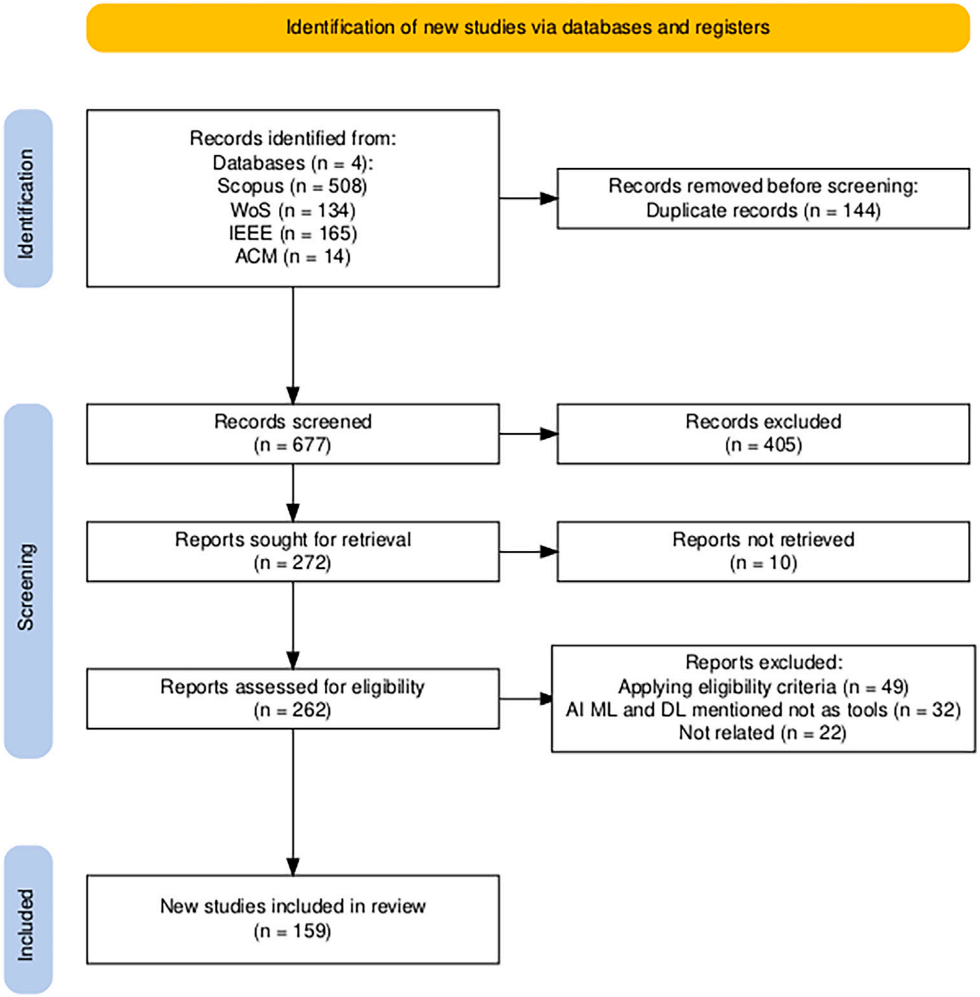

Identification of new studies via databases and registers
Identification
Records identified from:
Databases (n = 4):
Scopus (n = 508)
WoS (n = 134)
IEEE (n = 165)
ACM (n = 14)
Records removed before screening:
Duplicate records (n = 144)
Screening
Records screened
(n = 677)
Records excluded
(n = 405)
Reports sought for retrieval
(n = 272)
Reports not retrieved
(n = 10)
Reports assessed for eligibility
(n = 262)
Reports excluded:
Applying eligibility criteria (n = 49)
AI ML and DL mentioned not as tools (n = 32)
Not related (n = 22)
Included
New studies included in review
(n = 159)

Fig. 1. PRISMA 2020 flow diagram illustrating the study selection process, including database identification, screening, eligibility assessment, and final inclusion of studies in the SLR.

Finally, any limitations and suggestions for future work were recorded. This information not only provided a balanced perspective on the strengths and weaknesses of each study but also pointed to possible directions for further investigation and improvements in the field. By systematically extracting and documenting these elements, this structured approach enabled a comprehensive synthesis of the current state of knowledge and supported the identification of gaps and challenges in AI-driven formative and summative assessment research.

2.2.6.2. Data synthesis and analysis. The data synthesis process was designed to systematically analyze and consolidate the findings from the included studies. The results were presented using tables, figures, and network visualizations to analyze trends in publication years, geographic distribution, academic disciplines, publishers, and AI techniques. This synthesized evidence formed the basis for addressing the research questions, highlighting limitations in the existing literature, and informing the overall conclusions of the review.

The distribution of studies across databases was analyzed to examine the relative contribution of each source, guiding future researchers seeking comprehensive and relevant resources. In addition, the time distribution of the studies was investigated to identify trends in research activity over time, revealing periods of increased scholarly attention to AI-based assessment in higher education.

Geographic distribution was analyzed to provide insight into the global research landscape and to identify regions that have contributed most prominently to the field, highlighting opportunities for international collaboration and knowledge exchange [135]. Publication venues were also examined by analyzing the distribution of studies across journals, conferences, and publishers, allowing the identification of key outlets that regularly disseminate influential research at the intersection of education and computer science.

The distribution of studies across academic disciplines was further examined to assess methodological and conceptual diversity within the literature. This analysis revealed interdisciplinary contributions and the integration of ideas and techniques across domains. By applying these quantitative and visual analyses, the synthesis provided a comprehensive overview of the current research landscape on AI-driven formative and summative assessment and its role in predicting student outcomes.

In addition to descriptive statistics, the synthesis focused on identifying cross-study patterns, methodological trends, and emerging gaps to support higher-level interpretation in the Results and Discussion sections.

2.2.7. Quality assessment

According to the suggestion of [50], which is to determine the quality evaluation questions that best fit a particular study topic, researchers should review the collection of questions within the context of their paper. Therefore, this suggestion is adopted, and the quality assessment criteria for the primary studies are defined as follows.

Table 6
Quality assessment scoring scheme.

Applicable tool	Yes responses (out of 5)	Quality level	Interpretation
MMAT / CASP	5	High	Meets all quality criteria; strong methodological rigor.
MMAT / CASP	3–4	Medium	Meets most criteria with identifiable methodological limitations.
MMAT	1–2	Low	Substantial methodological limitations; findings interpreted descriptively.

To analyze the design and analytical strengths and limitations of quantitative, qualitative, and mixed-methods studies, a detailed critical appraisal was performed on the chosen articles. Studies reported in [67] employed the Mixed Methods Appraisal Tool (MMAT) to perform a critical appraisal. The MMAT is a valuable tool for evaluating the quality and potential biases in quantitative and mixed-method research articles. It outlines five primary quality criteria to evaluate a study's research design, sampling approach, and the methods and techniques used for data collection and analysis. For qualitative research, the Critical Appraisal Skills Programme (CASP) checklists (CASP, 2019) were used to critically examine ethical aspects and transparency in research practices. Studies meeting all five quality criteria were classified as high quality, those meeting three or four criteria as medium quality, and those meeting one or two criteria as low quality. Quality ratings were used to inform interpretation rather than to exclude studies.

Each study was evaluated against five quality criteria using the appropriate appraisal tool based on its methodological design. Studies meeting all five criteria were classified as high quality, those meeting three or four criteria were classified as medium quality, and studies meeting one or two criteria were classified as low quality, as shown in Table 6.

As documented in the quality evaluation results in Supplementary Table A.9, Table A.10, and Table A.11, most quantitative and mixed-method studies demonstrated medium to high quality, while a small number of quantitative studies were classified as low quality. Common limitations identified across medium and low quality studies included insufficient reporting of sample representativeness and potential nonresponse bias, particularly due to small sample sizes or limited information on participant recruitment. Qualitative studies generally met high quality standards under the CASP tool, with medium quality ratings primarily reflecting gaps in ethical or privacy considerations.

Quality assessment outcomes were used to inform the interpretation and weighting of findings rather than to exclude studies, with greater emphasis placed on evidence derived from high and medium quality studies when concluding.

2.3. Reporting the SLR

Following the execution of the SLR protocol described in Sections 2.2.1–2.2.7, the reporting phase focused on transparently summarizing the outcomes of study selection, data extraction, synthesis, and quality assessment. Findings were presented using qualitative and quantitative analyses supported by tables, figures, and network visualizations to communicate patterns related to assessment types, AI techniques, and reported student outcomes.

3. Results

This section presents the outcomes of the SLR, including study selection results, quality assessment of the included studies, network visualizations, and synthesized findings addressing the research questions.

3.1. Search results

The initial database search yielded a total of 821 records, including 508 from Scopus, 134 from Web of Science, 165 from IEEE Xplore, and 14 from the ACM Digital Library. After removing 144 duplicate records, 677 studies remained for screening.

Titles, abstracts, and keywords were then reviewed, resulting in the exclusion of 405 studies and the retention of 272 articles for further examination. Full-text retrieval was attempted for these articles, of which 262 were successfully assessed, while 10 records were excluded due to the unavailability of full text.

Following full-text assessment and application of the inclusion and exclusion criteria, 103 studies were excluded for not meeting the eligibility requirements. Ultimately, a total of 159 studies were included in the final review. The complete study selection process is illustrated in Fig. 1, in accordance with the PRISMA 2020 guidelines.

3.2. SLRs quality evaluation

Among the 112 quantitative studies analyzed (70.4%), 44 studies (27.7%) demonstrated high quality, achieving MMAT scores of 100%, which indicates complete adherence to all quality criteria. Sixty-one studies (38.3%) were classified as medium quality, with MMAT scores ranging from 60% to 80%, reflecting partial adherence to the criteria. Seven studies (4.4%) were classified as low quality, with MMAT scores between 20% and 40%, indicating limited compliance with the quality assessment criteria.

Similarly, among the 33 mixed-method studies (20.7%) analyzed, 15 studies (9.4%) demonstrated high quality with MMAT scores of 100%, indicating full adherence to the assessment criteria. The remaining 18 studies (11.3%) were classified as medium quality, with MMAT scores between 60% and 80%, reflecting partial fulfillment of the five MMAT criteria. No mixed-method studies were classified as low quality.

Notably, several quantitative and mixed-method studies did not fully satisfy MMAT criteria 1.2 (sample representativeness) and 1.4 (risk of nonresponse bias). A recurring limitation was the relatively small or narrowly defined sample sizes, which restrict the representativeness of the target population. For example, studies reported in [15,82,84,138,151] relied on limited datasets, making it difficult to determine whether the participants adequately represented the broader student population. In another instance, the study reported in [42] evaluated only five essay samples corresponding to five students, which may contribute to biased assessment outcomes.

Additionally, several studies [21,59,124,137] did not clearly report participant recruitment details, including the number of students invited and those who ultimately participated, making it challenging to assess potential nonresponse bias. If non-participating students systematically differed from participants, the reliability and generalizability of the findings might be affected. Detailed MMAT results for quantitative and mixed-method studies are presented in Supplementary Table A.9 and Table A.10.

Within the qualitative studies (8.8%, n = 14), eight studies met all CASP quality criteria and were therefore classified as high quality. The remaining qualitative studies were classified as medium quality, with common shortcomings related to ethical considerations and privacy reporting when presenting key findings. A detailed breakdown of the CASP-based quality assessment for qualitative studies is reported in Supplementary Table A.11.

Overall, these quality assessment results suggest that the majority of evidence in the reviewed literature is methodologically robust, while also highlighting recurring design limitations that should be considered when interpreting findings and designing future AI-based assessment studies. Identified methodological limitations, particularly related to sample representativeness and nonresponse bias, were common across medium- and low-quality studies and were considered when interpreting synthesized findings.

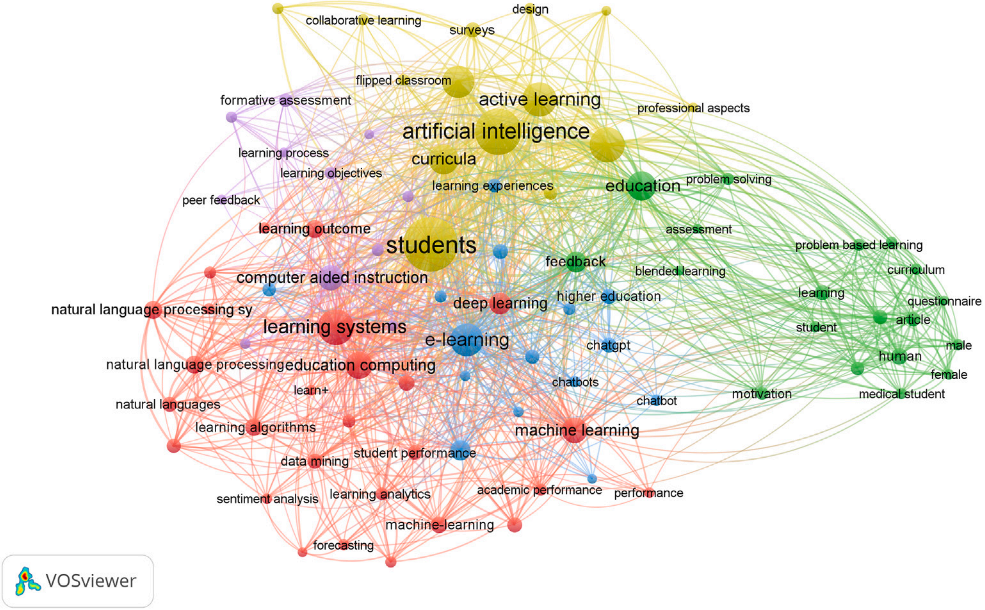


Fig. 2. Keyword co-occurrence network in AI-driven assessment and feedback studies, illustrating the convergence of student-centered learning, AI techniques, and assessment practices in higher education.

3.3. Network visualizations

3.3.1. Network visualization for keywords co-occurrence

Fig. 2 presents a keyword co-occurrence network for AI-based assessment and feedback studies, generated using Scopus data and visualized with VOSviewer. Each node represents a keyword, with node size indicating term frequency and link thickness reflecting co-occurrence strength. The terms “students” and “artificial intelligence” appear as the most dominant and highly connected keywords, highlighting the strong emphasis on student-centered applications of AI in higher education assessment research.

The network reveals five main keyword clusters, reflecting distinct yet interconnected research themes. Overall, the clustering structure indicates that AI-driven assessment research is shaped by the interaction between pedagogical priorities, technical methods, and learning environments rather than isolated methodological or educational perspectives.

- The yellow cluster emphasizes student-centered learning approaches, including concepts such as curricula, collaborative learning, flipped classrooms, and active learning, suggesting that AI-driven assessment is frequently aligned with pedagogical strategies aimed at improving student engagement and learning outcomes.
- The red cluster concentrates on technical foundations, including ML, DL, NLP, and data mining, reflecting the dominant methodological tools supporting AI-based assessment systems.
- The green cluster focuses on educational practices such as assessment, feedback, motivation, and blended learning, highlighting the role of AI in supporting instructional decision-making and learning support.
- The purple cluster relates to instructional design and formative assessment, including learning objectives and computer-aided instruction, indicating a strong connection between AI-based tools and formative assessment processes.
- The blue cluster represents higher education and e-learning technologies, including chatbots and ChatGPT, reflecting the growing interest in conversational and generative AI tools within educational assessment contexts.

Overall, the keyword network demonstrates that AI-based assessment research in higher education is inherently interdisciplinary, integrating pedagogical goals with advanced AI techniques. The prominence of student- and feedback-related terms further indicates that current research largely prioritizes formative assessment and learning support rather than purely summative or grading-focused applications.

Taken together, these clusters indicate that AI-driven assessment research in higher education is primarily oriented toward pedagogically grounded, student-centered applications, with technical AI methods serving as enablers rather than standalone research objectives.

3.3.2. Network overlay visualization for keywords co-occurrence

Fig. 3 presents a keyword overlay visualization that reflects the temporal development of research themes in AI-based assessment studies. The color range represents the average publication year of keywords, with earlier research trends shown in darker tones and more recent themes appearing in lighter colors.

The visualization indicates that early studies primarily focused on traditional AI techniques, such as ML, data mining, and learning analytics, applied to student performance and learning systems. In contrast, keywords associated with LLMs, chatbots, and ChatGPT appear predominantly after 2022, demonstrating a recent shift in research attention toward generative and conversational AI applications in educational assessment.

This temporal pattern suggests that while AI-based assessment research has progressively evolved over time, the rapid growth of LLM-related studies after 2022 reflects increasing interest in automated feedback, textual assessment, and interactive learning support. The overlay visualization therefore highlights a transition from predictive and analytic assessment approaches toward more generative and language-centered assessment applications in higher education.

3.3.3. Network visualization of co-authorship among nations

Fig. 4 illustrates the international co-authorship patterns among countries contributing to AI-based assessment and feedback research in higher education. In this network, each node represents a country, with node size indicating publication volume, while links denote collaborative co-authorship relationships.

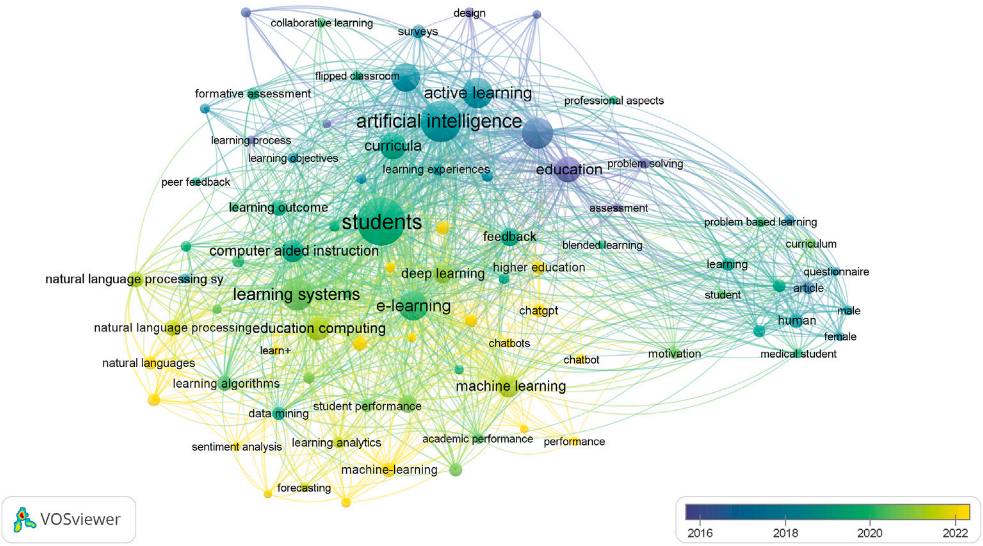


Fig. 3. Keyword overlay visualization showing the temporal evolution of AI-driven assessment research, highlighting the recent emergence of LLM-related terms in higher education.

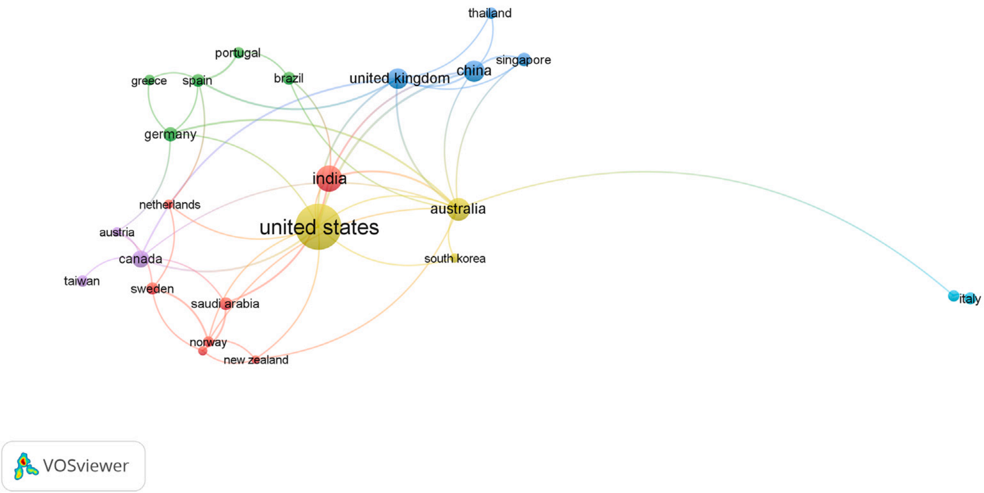


Fig. 4. Network representation of international co-authorship patterns in AI-based assessment and feedback research in higher education.

Fig. 4 shows co-authorship relationships among more than 20 countries, indicating a moderate level of international collaboration in this research area. The United States emerges as the most prominent contributor, followed by India and Australia, reflecting their strong research capacity and active engagement in AI-driven educational assessment studies. In terms of collaboration, Australia shows the highest total link strength, followed by the United States and the United Kingdom, suggesting that these countries play a central role in facilitating cross-national research partnerships.

Overall, the co-authorship network indicates that while AI-based assessment research is geographically diverse, collaboration tends to be concentrated among a small number of leading countries. This pattern suggests that research output and international collaboration are influenced by established research infrastructures, funding availability, and institutional support. Expanding cross-regional collaboration may help promote more balanced global contributions and knowledge exchange in AI-driven assessment research.

These collaboration patterns suggest that AI-based assessment research is shaped not only by technological readiness but also by national research investments and institutional capacity, potentially reinforcing inequalities in global research participation.

3.4. Answering research questions

This section aims to answer the research questions of the study, as detailed in Table 4:

3.4.1. What is the current state of research on the implications of AI-driven formative and summative assessments and feedback for student outcomes in higher education?

Assessment and feedback mechanisms in higher education significantly influence student learning outcomes and educational effectiveness. Recently, scholarly attention has increasingly focused on exploring the roles of formative and summative assessments, particularly within the context of advanced technologies such as AI techniques, including ML, DL, and LLMs. This question synthesizes the current state of research on AI-driven formative and summative assessments and feedback, focusing on publication trends, geographical distribution, assessment implications, and disciplinary contexts in higher education. The following subsections present a detailed analysis of these trends.

3.4.1.1. Trends in AI-based assessment research. The distribution of the included studies on AI-driven assessments in higher education over time

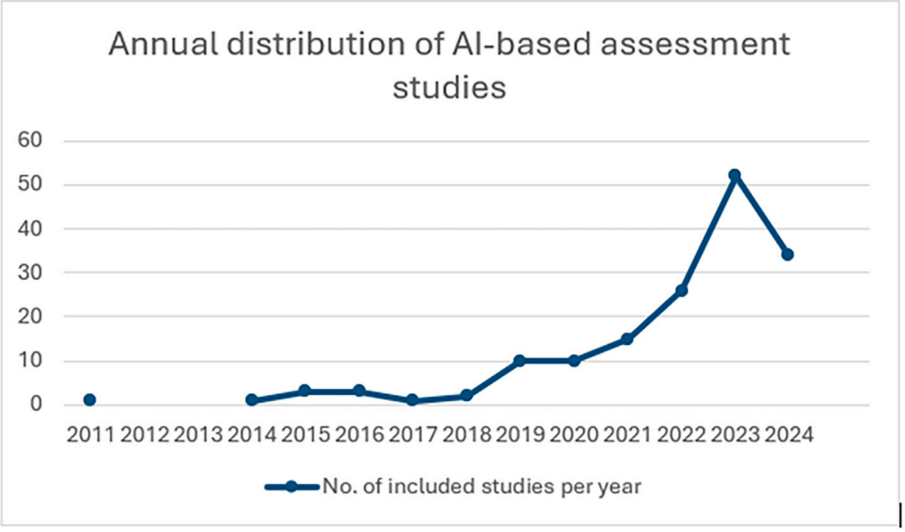


Fig. 5. Annual growth of AI-based assessment research in higher education, showing a sharp increase after 2019 and a peak in 2023.

is presented in Fig. 5. Research activity remains relatively limited before 2019, followed by a noticeable and continuous increase in more recent years. In particular, a sharp growth is observed from 2021 onwards, with the highest number of publications reported in 2023. This increasing trend suggests a growing research interest in AI-based assessment within higher education, which may be associated with recent advances in AI technologies and the increased adoption of digital assessment systems in educational institutions.

Table 7
Regional distribution of AI-based assessment studies.

Region	Percentage of studies
Asia	45.68%
North America	22.22%
Europe	21.60%
Oceania	4.94%
South America	3.70%
Africa	1.85%

3.4.1.2. Geographical distribution of research. Table 7 presents the geographical distribution of AI-based assessment studies in higher education across 45 countries. The results show clear regional variation, with Asia contributing the largest share of studies, followed by North America and Europe. Within Asia, China represents the most prominent contributor, while research activity in North America is largely driven by the United States. European research output is more evenly distributed across multiple countries. In contrast, relatively limited research activity is observed in South America, Africa, and Oceania, with contributions concentrated in a small number of countries.

The dominance of studies from China and the United States may be related to broader patterns in global AI research output and investment. Bibliometric evidence indicates that China and the United States collectively lead AI publication volume and citation impact, reflecting substantial national investment, research capacity, and well-established AI research infrastructures [13]. These conditions likely facilitate greater research activity in specialized domains such as AI-driven educational assessment. Conversely, lower representation from other regions may reflect disparities in research funding, technological readiness, and data accessibility rather than a lack of scholarly interest in AI-based assessment.

3.4.1.3. Publication trends. The included studies on AI-based assessment in higher education were published across both conference proceedings and journal venues. Overall, conference papers constitute a slightly larger proportion of the included studies compared to journal articles, as illustrated in Supplementary Figure A.9. This highlights the prominent role of conferences in disseminating research in this area.

In terms of publishers, the distribution indicates a strong presence of conference-oriented venues, particularly those associated with computer science and engineering disciplines. The dominance of conference publications may reflect the fast-moving nature of AI research, including ML, DL, and LLM research, in which preliminary findings and emerging methods are often shared at conferences before they are published in extended journal versions. At the same time, journal publications are distributed across a wide range of outlets, highlighting the interdisciplinary character of AI-based assessment research across education, technology, and engineering domains. Detailed information on journal-level publication distribution is provided in Supplementary Table A.12.

Overall, the prominence of conference venues reflects the rapid methodological evolution of AI-based assessment research, while the growing presence of journal publications suggests increasing consolidation, validation, and interdisciplinary engagement within the field.

3.4.1.4. Formative, summative, and both assessment implications. The included studies were classified into three main assessment categories: formative, summative, and both. Studies categorized as using both assessments refer to cases in which AI-based techniques were applied to support formative and summative assessment purposes within the same study. In these cases, studies reported the use of AI-based techniques to support both formative feedback and summative evaluation, either through a single AI model or through multiple AI-based tools applied within the same educational context. However, in many studies, it was not explicitly reported whether the use of both formative and summative assessment purposes was supported by a single AI system or by multiple AI-based tools applied at different stages of the assessment process, as this information was not consistently reported across studies. These cases reflect study-specific integration of assessment purposes rather than systematic integration across the broader literature.

As shown in Fig. 6, formative assessment represents the dominant category across the reviewed studies, while summative-focused applications remain comparatively limited. Within formative assessment, most studies emphasize feedback, learning progress, and student engagement, whereas summative assessment primarily focuses on performance evaluation. Studies applying both formative and summative approaches

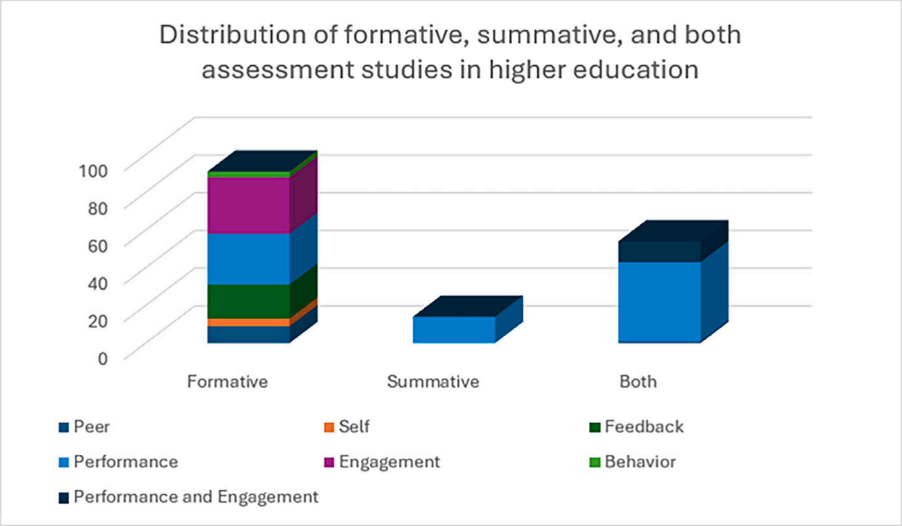


Fig. 6. Distribution of formative, summative, and both assessment types in AI-based studies, showing the dominance of formative assessment and performance-focused applications.

also show a strong emphasis on performance-related outcomes, with relatively limited adoption of peer- or self-assessment strategies.

Overall, these findings indicate that AI-based assessment research in higher education is primarily oriented toward supporting learning processes through formative assessment and feedback rather than high-stakes summative evaluation. The comparatively low representation of summative and peer- or self-assessment approaches highlights ongoing challenges related to reliability, validity, and trust in automated assessment systems, particularly in evaluative contexts.

3.4.1.5. Discipline research trends. Each included study was examined to identify the academic discipline from which student data were sourced, as summarized in Supplementary Figure A.10. The results indicate a strong concentration of AI-based assessment research within STEM-related disciplines, with computer science representing the most dominant domain. Engineering, science, and health-related fields also show notable representation, reflecting the alignment between computational methods, data availability, and assessment applications in these domains.

In contrast, humanities, social sciences, business, and education-related disciplines are less frequently represented. Within these areas, language-focused studies constitute the largest proportion of contributions, while disciplines such as arts, business, and education appear only occasionally. A limited number of studies adopt multidisciplinary perspectives by combining data across domains or applying AI-based assessment methods in diverse learning contexts.

Notably, a substantial subset of studies does not explicitly report the disciplinary context of the data used, indicating inconsistencies in reporting practices and highlighting the need for clearer documentation of application domains in AI-driven assessment research. Overall, these findings suggest that AI-based assessment has primarily evolved within technically oriented disciplines, where automated evaluation and performance-based tasks are common, while presenting opportunities for future research to adapt AI-driven assessment methods to a broader range of pedagogical contexts and learning outcomes.

3.4.2. What AI techniques, including ML, DL, and LLM, are used for formative and summative assessments and feedback to enhance student outcomes?

To address this research question, the reviewed studies were analyzed by categorizing AI techniques according to their primary learning paradigms and assessment functions. Rather than functioning as isolated methods, these techniques exhibit complementary strengths and limitations depending on assessment objectives, data type, and learning context. Supplementary Table A.13 summarizes the distribution of AI techniques across formative, summative, and both assessment contexts, highlighting the dominance of classification-based approaches and the emerging role of generative models.

The integration of AI, ML, DL, and LLM techniques in educational assessments has gained significant attention. Recent studies increasingly distinguish LLM-based approaches from traditional NLP techniques used in earlier assessment systems. As reflected in the distribution of generative and text-mining techniques identified in this review, LLMs differ from rule-based or feature-engineering-driven NLP methods by enabling more flexible processing of student-generated text and supporting automated feedback generation, tutoring, and conversational assessment. However, several challenges remain evident in the reviewed literature, particularly in relation to hallucination risks, response consistency, and reliability when LLMs are applied in high-stakes summative assessment contexts. Consistent with the observed concentration of summative studies and dataset limitations, these issues raise concerns regarding assessment validity, grading fairness, and authorship verification. As a result, most LLM-based applications identified in this review are predominantly deployed in formative assessment settings, where feedback and learning support are emphasized rather than final grading decisions.

Table 8 provides a comparative analysis of the identified AI techniques, summarizing their typical assessment purposes, data types, student outcomes, strengths, and limitations, while the following subsections discuss each technique in more detail.

3.4.2.1. Classification. The most popular technique in higher education is classification, which is a supervised learning approach designed to predict class labels based on input features. Classification techniques have been widely applied to enhance student outcomes by supporting diverse assessment tasks, including evaluating student programming skills to improve course delivery and identify struggling students [21,27,156], assessing student responses through automated systems to provide personalized feedback [35,69,79], and analyzing learning behaviors, engagement, and interaction to enhance teaching practices and learning effectiveness [15,18,61,73,82,134]. In addition, classification has been used to assess student feedback quality [1,19,108,145,152], verify student identity in oral and online assessments using audio responses [115],

Table 8
Comparative analysis of AI techniques used in formative and summative assessments.

AI technique	Assessment purposes	Typical data types	Typical student outcomes	Strengths	Limitations	Representative studies
Classification	Performance prediction, early warning, feedback automation, engagement analysis	Structured logs, text, speech	Performance, engagement, feedback	High predictive accuracy; interpretable decision support; effective for early intervention	Sensitive to class imbalance; depends on labeled data	Section 3.4.2.1
Regression	Score prediction, grade estimation, risk modeling	Numerical and temporal data	Scores, grades	Quantifies continuous outcomes; useful for progression tracking	Limited for categorical feedback; assumes linearity	Section 3.4.2.2
Clustering	Learner profiling, behavior discovery, engagement grouping	Interaction logs, activity traces	Learning behavior, engagement	Uncovers latent patterns; no labels required	Results harder to interpret; not predictive	Section 3.4.2.3
Combination Approaches	Comprehensive outcome prediction and learner modeling	Mixed data	Performance, engagement, feedback	Captures complementary strengths of techniques	Higher complexity; reduced interpretability	Section 3.4.2.4
Generative Data	Automated feedback, text evaluation, conversational assessment	Essays, discussions, chats	Feedback, writing, coding	Handles open-text responses; scalable formative feedback	Hallucination risk; reliability issues in summative use	Section 3.4.2.5
Text Mining	Sentiment analysis, feedback classification	Textual data	Feedback, literacy	Efficient content analysis; interpretable features	Limited contextual understanding compared to LLMs	Section 3.4.2.6
Other Techniques	Specialized assessment tasks (e.g., speech, simulation)	Domain-specific data	Varies	High task specificity	Limited generalizability	Section 3.4.2.7

appraise speeches to support language learning [29,77], and evaluate electronic virtual patient systems for healthcare education [20].

In supervised classification, algorithms such as random forest [2,4–6,63,72,123], convolutional and artificial neural networks [45,68,87,133], support vector machines [7,19,45,121], logistic regression [36,41,89], and NLP-based models such as BERT [21,35,86] have been extensively utilized due to their robust predictive capabilities. Consequently, classification approaches enable flexible assessment across different contexts and support the early identification of students who may require timely academic assistance, allowing decision-makers to provide proactive interventions and improve academic performance.

3.4.2.2. Regression. Regression techniques are used to identify relationships between dependent and independent variables and are primarily applied to predict continuous and numerical outcomes. Unlike classification, which predicts discrete classes, regression estimates values such as grades or scores. In higher education assessment, a limited number of studies have employed regression to predict student performance and scores to support peer assessment, feedback processes, and teaching practices [7,65,83,141]. Common regression approaches include linear and Bayesian regression models [60,66,105], as well as random forest-based regression [64,78].

Overall, regression techniques have been used to forecast student performance and identify factors associated with academic risk in tertiary education. However, the relatively small number of studies highlights a potential gap in the literature, suggesting opportunities for further research on regression-based assessment approaches in higher education.

3.4.2.3. Clustering. Clustering is an unsupervised learning technique that groups similar data instances to reveal underlying patterns in educational data. In higher education, clustering has been applied to enhance exam pass rates [98], analyze student performance [33,34], assess language learning and writing activities [114,139], and examine student reflections, engagement, and feedback processes [113,126,128]. Through these applications, clustering supports the identification of learning patterns that can inform teaching practices, curriculum adjustments, and early warnings of performance changes.

Most studies employed K-means clustering [32,110,137] due to its effectiveness in exploratory analysis without predefined labels. Similar to regression, the limited adoption of clustering techniques highlights a gap in current assessment research and suggests potential for expanded use in supporting data-driven decision-making in higher education.

3.4.2.4. Combination approaches. The advancement of AI techniques has enabled the use of combination approaches that integrate classification, regression, and clustering to analyze complex educational data. Several studies have combined classification and regression to assess student performance, feedback, public speaking skills, and active learning outcomes [12,46,100,154]. Other studies have integrated classification with clustering to identify at-risk students, grade responses, and assess student answers [84,99,140]. In addition, regression and clustering have been combined to predict final exam scores based on student engagement patterns [52].

These combination approaches allow for a more comprehensive analysis of student performance and engagement, supporting informed decision-making and timely interventions in higher education assessments.

3.4.2.5. Generative data. Generative AI techniques focus on creating new content, particularly human-like text, when applied through LLMs. In higher education, these techniques have been integrated with ML, DL, and NLP approaches to assess students' creative work, programming skills, language learning, argumentation, feedback, engagement, and cognitive levels [47,48,94,97,118,122,137,144]. The majority of generative assessment approaches rely on GPT-based LLMs [11,23,38,132,137] to support adaptive feedback and interactive learning environments. A small number of studies have applied generative models within supervised learning frameworks, typically using labeled datasets to evaluate generated responses or feedback quality.

The reviewed studies indicate a growing adoption of generative and LLM-based approaches in higher education assessment, particularly for tasks involving automated feedback, student response evaluation, programming assessment, and analysis of written text. Supplementary Table A.14 shows that most LLM-based applications rely on general-purpose models, such as GPT-based architectures, primarily deployed in formative assessment settings to support feedback generation, engagement analysis, and learning support. In contrast, relatively few studies report task-specific adaptation or fine-tuning of LLMs tailored to particular assessment objectives. This reliance on generic models highlights methodological limitations related to reliability, consistency of outputs, and contextual sensitivity, especially when processing informal or diverse student-generated content [106,127]. As reflected in the reviewed studies, these limitations restrict the use of LLMs in high-stakes summative assessment contexts, where grading accuracy, authorship verification, and assessment validity are critical. Overall, the findings suggest that while LLMs demonstrate strong potential for enhancing formative assessment practices, further research is needed to explore customized and task-adapted LLM approaches to improve robustness and applicability in summative assessment scenarios.

These findings suggest that generative techniques can enhance assessment flexibility and feedback quality; however, concerns related to ethical considerations, privacy, reliability, and appropriate use in high-stakes assessments remain important challenges for future research.

3.4.2.6. Text mining. Text mining techniques analyze unstructured textual data to extract meaningful information in educational contexts. In higher education, text mining has been applied to enhance student engagement, provide constructive feedback, identify early indicators of performance, and assess digital literacy skills [14,40,112,116]. NLP-based models are commonly employed to support semantic analysis and feedback generation [14,40,70].

These techniques enable the analysis of student submissions, online discussions, and assessment content, contributing to improved feedback processes, reduced grading bias, and enhanced assessment validity and reliability.

3.4.2.7. Other techniques. Other AI-related techniques include dimensionality reduction, decision-making models, statistical analysis, information retrieval, chatbots, fuzzy logic, and reinforcement learning. In higher education, these techniques have been applied to assess feedback quality and satisfaction [39,85,119], enhance interactive learning environments [10,55,147], evaluate language learning outcomes [81], support real-time feedback systems [31,53], and automate grading processes [42,105].

These approaches are often integrated into larger assessment systems to monitor student behavior, automate evaluation, and improve the overall quality of assessment and feedback in higher education.

3.4.3. Is there any existing evidence regarding the effectiveness of AI-driven formative and summative assessment applications in predicting student outcomes?

Despite the use of formative and summative assessments in enhancing student outcomes, this research question elaborates on the existing evidence regarding the effectiveness of formative and summative assessments in predicting student outcomes in higher education. Overall, assessments were frequently employed for the prediction of student outcomes, indicating a strong research interest in this direction. The included educational assessment studies were categorized into three groups: formative assessment, summative assessment, and formative and summative assessment (both). Together, these studies provide empirical evidence that AI-supported assessment, particularly formative and both approaches, can effectively support early prediction, intervention, and learning improvement in higher education.

The findings show that 64 out of the 159 included studies explicitly focused on predicting student outcomes using AI-based assessment approaches. Among these studies, formative assessment and both assessment approaches were more commonly applied than summative assessment alone (see Fig. 7). In particular, studies applying both assessment types represented the largest proportion, suggesting that integrating assessment information across different stages of learning is a common strategy for outcome prediction.

Fig. 7 also illustrates how prediction functions are achieved in educational assessment through the distribution of AI techniques employed. Most of the reviewed studies relied on ML techniques, either alone or in combination with other approaches such as NLP or DL. This indicates that ML remains the dominant predictive technique in educational assessment, likely due to its effectiveness in handling structured educational data and extracting predictive patterns related to student learning. A smaller number of studies employed DL methods or combined ML and DL techniques, while techniques such as NLP, explainable AI, and generative AI were used less frequently.

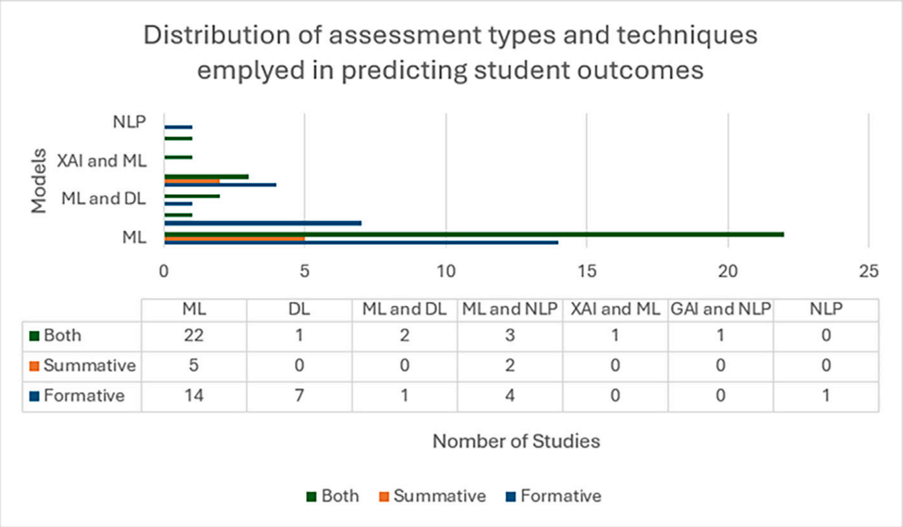
In terms of predicted student outcomes, academic performance was the most commonly investigated outcome, followed by student engagement (Fig. 8). Fewer studies focused on predicting student scores, behaviors, emotional states, sentiment, or the quality of student reviews. These findings suggest that performance and engagement are considered the most relevant and measurable indicators of student outcomes in AI-based assessment research.

When examining the relationship between assessment types and predicted outcomes, formative assessment was primarily used to predict engagement, behavior, and emotional or sentiment-related outcomes. In contrast, summative assessment was more commonly associated with predicting final performance or scores. Studies applying both formative and summative assessments tended to focus on performance prediction, indicating that combining ongoing assessment data with final evaluation results can enhance prediction accuracy (see Fig. 8). This pattern suggests that formative assessment is particularly effective for modeling dynamic learning processes, while summative assessment remains primarily aligned with outcome certification rather than early prediction.

Evidence from individual studies further supports these observations. For example, in the formative assessment category, the work reported in [4] developed an explainable ML-based approach to predict student performance at the assignment and quiz levels. The results showed that random forest models achieved high predictive accuracy while also providing interpretable insights that supported feedback and learning improvement. In summative assessment contexts, the study described in [104] introduced an ML-driven automatic assessment system for evaluating 3D computer animation coursework, demonstrating more consistent scoring compared to manual grading. Moreover, studies combining both assessments, such as [75], showed that integrating data from automatic marking systems, discussion forums, and assessment scores enabled early and accurate prediction of students' exam results.

Overall, the findings indicate that there is substantial evidence supporting the effectiveness of AI-based formative assessment and both assessment approaches in predicting student outcomes. In contrast, summative assessment is less frequently used for prediction purposes, possibly due to its high-stakes nature and limited data availability. These results highlight the importance of formative and both assessment strategies for supporting prediction, early intervention, and improved student outcomes in higher education. These findings align with learning analytics and formative assessment theories, which emphasize continuous feedback and data-driven insights as key enablers of student success.

3.4.3.1. Common features used. AI-based assessments commonly utilize a range of student-related features to predict outcomes such as academic performance, scores, emotions and sentiments, engagement, behavior, and the quality of student reviews, as summarized in Supplementary Table A.14. These features include quiz results, assessment scores, frequency and consistency of task completion, and syllabus coverage. Historical academic performance is frequently employed, reflecting a strong reliance on prior learning behavior as an indicator of future outcomes. Additionally, clickstream activity logs and demographic information (e.g., age or gender) are often used to identify students at risk of failing or dropping out. Alongside these quantitative indicators, textual and narrative feedback from instructors or peers provides qualitative insights that enhance predictive accuracy. Engagement-related metrics consistently emerge as influential predictors across multiple studies.



	ML	DL	ML and DL	ML and NLP	XAI and ML	GAI and NLP	NLP
🟩 Both	22	1	2	3	1	1	0
🟧 Summative	5	0	0	2	0	0	0
🟦 Formative	14	7	1	4	0	0	1

Fig. 7. Distribution of assessment types and AI techniques employed in studies predicting student outcomes, showing the dominance of ML and both assessment across formative, summative, and integrated assessment designs.

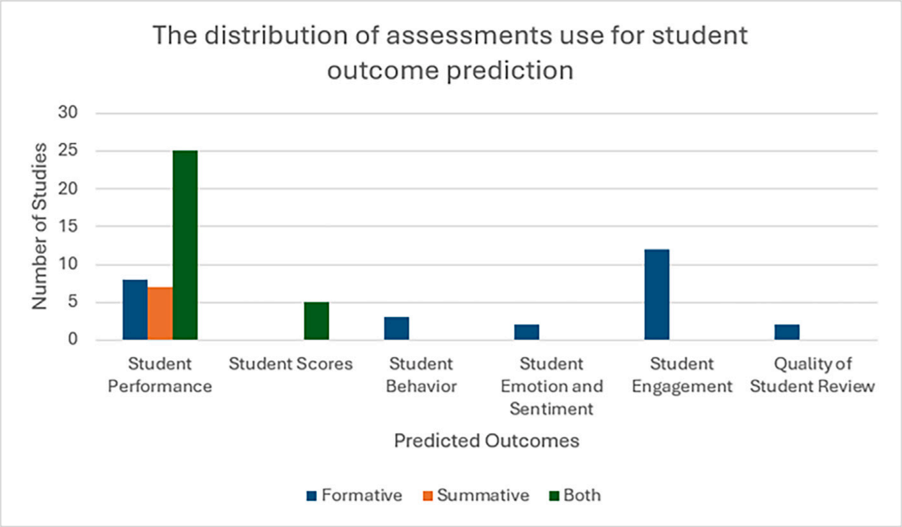


Fig. 8. Distribution of predicted student outcomes across formative, summative, and both assessment types, highlighting the predominance of performance and engagement prediction in AI-based assessment research.

For example, the study reported in [83] collected feedback from two consecutive assignments in a postgraduate data science course, using grades from the first assignment as a baseline to capture changes in student performance. Features such as initial assignment grades, use of interrogative language, and positive emotional expressions were found to be significant predictors of grade improvement in subsequent assignments.

Similarly, research described in [65] trained predictive models using grades assigned by both peers and instructors, explicitly accounting for assessor bias and grading precision. Instructor grades were used as ground truth, enabling the prediction of final grades while improving the validity of peer-assessed outcomes.

In another study, [157] combined quiz scores, syllabus coverage rates, the number of completed quizzes, student feedback, and historical performance records within the QLearn platform. These features were analyzed using ML algorithms to predict student progress, identify areas requiring improvement, and estimate final examination performance, supporting adaptive learning and personalized intervention.

Finally, the authors of [75] integrated data from an automatic marking and feedback system (PASTA), a discussion forum (Piazza), and assessment scores collected throughout the semester. Features extracted from assessment marks and student activity were used to train a decision tree classifier, which achieved prediction accuracies of 72.69% at the end of the semester and 66.52% mid-semester. These results demonstrate how feature-rich models can support the early identification of student performance trends and timely academic intervention in higher education.

Across the reviewed studies, engagement- and performance-related features consistently emerge as the most influential predictors, suggesting that combining behavioral traces with assessment artifacts enables more robust and timely predictions of student outcomes.

3.4.4. What data sources are utilized for AI-driven assessments in higher education?

Various data sources are utilized for AI-based assessments in higher education. Overall, the datasets used in the reviewed studies can be classified into customized, public, and fictional datasets. The majority of studies relied on customized datasets ($n = 131$; 82.39%), followed by public datasets ($n = 19$; 11.95%). Among the customized datasets, 119 studies (74.84%) reported dataset sizes, while 12 studies (7.55%) mentioned the datasets without specifying size. For public datasets, 17 studies (10.69%) reported dataset sizes, and two studies (1.26%) did not. In addition, seven studies (4.40%) did not clearly specify the dataset used, and two studies (1.26%) employed fictional datasets, with one reporting dataset size and one not. These distributions are summarized in Supplementary Table A.15 and visually illustrated in Supplementary Figure A.11. The following categories summarize the primary dataset types identified in the reviewed studies:

1. **Customized Datasets**

   Customized datasets consist of private data collected directly from educational contexts and represent the most frequently used data source across the reviewed studies. These datasets can be further categorized as follows.

   Educational Assessment Data

   This category includes assessment-related measures such as test and assignment responses, scores [32,63,71,76,80], feedback [19,85,88,136], survey results [84,116,118,153], and other assessment artifacts [65,112,137] used to evaluate student understanding and the effectiveness of educational interventions. Historical academic performance and demographic information were also included in some studies [157]. This category is the most represented, with 58 studies, and dataset sizes ranging from small groups (e.g., 12 students) to large-scale datasets (e.g., over 12,000 MOOC participants).

   Educational Interaction Data

   This classification focuses on data capturing student interaction with learning systems, including log data [3,4,53,98,129], instructional modes, and video or audio recordings [5,41,45,77,82,115,121,133]. It also encompasses interaction-based grades, student responses, feedback, and surveys. In total, 49 studies employed interaction data to analyze engagement, emotional states, and learning behaviors, with dataset sizes varying substantially—from small classroom samples (e.g., 19 students) to large-scale datasets (e.g., millions of recorded interactions).

   Student Responses and Performance Data

   This category includes data directly related to students' responses and performance in quizzes, assignments, and exams, as well as engagement measures during learning activities [14,29,33,42,59,69,79,139]. Unlike interaction data, this category excludes system logs and survey data and focuses on individual-level performance metrics. Nineteen studies fall into this category, with dataset sizes ranging from small samples to several thousand students.

   Student Academic Records

   This category primarily includes final grades and academic performance records, occasionally linked to course-related employment outcomes [155]. Only one study relied on this data type, and the dataset size was not specified.

   Specialized Datasets

   A small subset of studies employed highly specialized datasets tailored to specific application domains, such as patient histories [20], physical activity data [15], Java code repositories [21], and neurosurgical performance metrics [49]. These datasets illustrate the diversity of customized data used in AI-based educational assessment, particularly in domain-specific training contexts.

2. **Public Datasets**

   Public datasets consist of openly accessible data obtained from external sources, as detailed in Supplementary Table A.15, which provides direct links to the datasets used. These datasets were less frequently employed than customized datasets but enabled reproducibility and cross-study comparisons in some cases.

3. **Fictional Datasets**

   Fictional datasets are artificially constructed for experimental purposes rather than being collected from real educational settings. Only two studies [9,126] employed fictional datasets, indicating that this approach is comparatively uncommon in AI-based assessment research in higher education.

The heavy reliance on customized, institution-specific datasets highlights persistent challenges related to data accessibility, reproducibility, and large-scale validation in AI-based assessment research, underscoring the need for greater use of shared benchmark datasets and transparent data reporting in future studies.

4. Discussion

This SLR investigated the impact of formative and summative assessments and feedback in higher education through the use of AI techniques, including ML, DL, and LLMs. Based on the synthesis of 159 primary studies, the findings provide robust empirical evidence on how AI-driven assessment practices support student learning, outcome prediction, and instructional decision-making. This section interprets the results through established educational and assessment theories, discusses practical implications, and highlights ethical, methodological, and policy-related challenges that shape the future development of AI-based assessment in higher education.

4.1. Theoretical interpretation of findings

4.1.1. Dominance of formative assessment and learning theory alignment

One of the most consistent findings of this review is the dominance of formative assessment in AI-driven educational research. The majority of studies apply AI techniques to support feedback, engagement monitoring, and early identification of at-risk students rather than high-stakes grading. From a pedagogical perspective, this trend aligns closely with constructivist learning theory, which emphasizes active knowledge construction through continuous interaction, feedback, and reflection. AI-driven formative assessment systems support these processes by generating timely, data-driven feedback and enabling students to regulate their learning paths.

The strong association between formative assessment and prediction tasks reflects principles of personalized learning, where assessment is used not only to measure achievement but to adapt instruction to individual needs. Predictive and adaptive models that use formative data allow instructors to intervene early, tailor support, and improve learning outcomes. In contrast, the relatively limited adoption of AI in summative assessment contexts reflects ongoing theoretical and practical concerns related to reliability, transparency, fairness, and accountability. Summative assessment serves certification and decision-making functions that require strong evidential standards, which many AI techniques, particularly generative models, do not yet consistently meet.

Overall, the findings reflect both the strengths and limitations of current AI-based assessment research. While significant progress has been made in developing formative assessment models, addressing gaps in summative, peer, and self-assessment approaches would contribute to a more comprehensive and balanced assessment framework. Expanding research across diverse educational systems and cultural contexts would further enhance the global relevance and equity of AI-based assessment practices.

Disciplinary patterns further contextualize these findings. The reviewed studies are heavily concentrated in STEM fields, particularly computer science and engineering, where structured assessment tasks and digital data are readily available and closely aligned with computational methods. In contrast, humanities, social sciences, and education-focused disciplines remain underrepresented, indicating substantial opportunities to extend AI-based assessment approaches to more diverse pedagogical contexts and learning outcomes.

4.1.2. Complementary roles of AI techniques across assessment purposes

The distribution of AI techniques identified in this review indicates that different methods are applied to address distinct assessment purposes, rather than being used interchangeably as universal solutions. Classification-based approaches dominate the literature due to their effectiveness in performance prediction, early warning systems, and feedback automation. These techniques align well with learning analytics frameworks that prioritize predictive insight and decision support for instructors and institutions.

Regression and clustering techniques, although less frequently applied, offer distinct contributions, which may partially explain their limited adoption in applied educational settings. Regression models support continuous outcome prediction, such as score estimation and student progression tracking, while clustering methods enable learner profiling and pattern discovery without predefined labels. Their limited use suggests not a lack of potential but rather greater methodological complexity and interpretability challenges within educational contexts. While several studies demonstrate promising applications, existing evidence remains preliminary, indicating the need for further research to optimize predictive accuracy, robustness, and pedagogical usefulness.

The emergence of generative AI and LLM-based approaches represents a conceptual expansion of AI-supported assessment. Unlike traditional predictive models, generative techniques enable language-centered assessment practices, including automated feedback, conversational assessment, and evaluation of open-text responses. The reviewed studies indicate that LLM-based applications are predominantly deployed in formative assessment settings, where feedback quality and learning support are prioritized. This pattern reflects persistent concerns related to hallucinations, response consistency, authorship verification, and transparency, which currently constrain the use of LLMs in high-stakes summative assessment contexts. Difficulties in processing informal language, mixed registers, and discipline-specific terminology further limit reliability in peer review and self-assessment scenarios [106,127].

Taken together, the findings suggest a functional division of labor among AI techniques: predictive models primarily support monitoring and early intervention, while generative models enhance feedback richness and learner engagement. Hybrid and integrated approaches combining multiple techniques appear promising for capturing complex learning dynamics; however, further research is required to optimize their predictive accuracy and adaptability across different assessment contexts.

4.1.3. Assessment and prediction of student outcomes

The strong emphasis on predicting student outcomes situates AI-based assessment solidly within the learning analytics approaches that emphasize prediction, early identification, and data-informed support. Most predictive applications draw on formative or both formative and summative assessment data, reinforcing theoretical perspectives that conceptualize assessment as an ongoing process for understanding learning paths rather than a retrospective evaluation of achievement. When summative data are combined with formative indicators, prediction accuracy improves, indicating that integrated assessment designs provide a more comprehensive representation of student learning behaviors.

Prediction effectiveness is closely linked to feature selection and data richness. Engagement metrics, historical academic performance, and interaction data consistently emerge as influential predictors, supporting theoretical assumptions that learning is shaped by behavior, effort, and context in addition to outcomes. These findings highlight the importance of longitudinal and multimodal data for advancing AI-driven assessment research in higher education.

4.2. Practical implications for higher education

From a practical perspective, the findings have clear implications for teaching practice, assessment design, and institutional policy. For instructors, AI-driven formative assessment systems can support timely feedback, early identification of learning difficulties, and targeted pedagogical interventions, reducing reliance on end-of-course examinations alone. AI-based tools should therefore be viewed as decision-support systems rather than replacements for instructor judgment.

The comparatively limited adoption of AI in summative assessment contexts suggests that high-stakes uses require cautious implementation. While AI can improve consistency and efficiency for structured evaluation tasks, hybrid human-AI assessment models remain the most appropriate approach for complex or subjective assessments. In such models, AI supports analysis and feedback generation, while educators retain responsibility for final grading and academic decisions.

At the curriculum level, the findings highlight the value of designing continuous, low-stakes assessments that generate meaningful data throughout the learning process. Integrating formative and summative assessment data enhances predictive accuracy and supports adaptive learning pathways. However, the heavy reliance on customized and institution-specific datasets underscores the need for standardized reporting and documentation practices to improve transferability and scalability.

4.3. Ethical, methodological, and policy challenges

Despite growing adoption, several ethical and methodological challenges persist. Many AI-driven assessment systems rely on historical and institution-specific data, which may encode existing inequities related to access, demographics, or curriculum design. The predominance of private datasets raises concerns regarding generalizability, reproducibility, and fairness.

Transparency and explainability remain central challenges. Although predictive models are frequently used to identify at-risk students, relatively few studies clearly explain model reasoning or feature contributions. This lack of interpretability affects trust and accountability, particularly when assessment results influence progression or certification.

The adoption of LLM-based assessment introduces additional risks, including hallucinations, response inconsistency, and sensitivity to informal language (e.g., slang, mixed linguistic registers, and discipline-specific terminology). The widespread reliance on general-purpose LLMs, rather than task-adapted models, limits robustness in formative contexts such as peer and self-assessment, as well as in summative evaluation scenarios. While these limitations primarily affect feedback quality and interpretability in formative assessment, they pose more substantial concerns in summative assessment settings, where reliability, transparency, and grading validity are critical. These challenges highlight the need for rigorous validation protocols, task-specific adaptation, and sustained human oversight before LLM-based systems are extended beyond formative contexts and considered for high-stakes summative assessment use.

These findings also highlight the need to balance context-specific (customized) datasets with greater use of public and well-documented data sources in order to enhance transparency, reproducibility, and cross-study comparability in AI-based educational assessment research. At the policy level, the review further reveals a lack of consistent governance frameworks guiding the use of AI in educational assessment. Institutions therefore need to establish clear policies addressing data privacy, informed consent, transparency, bias mitigation, and human oversight, alongside professional development initiatives that enable educators to interpret and responsibly apply AI-generated insights.

4.4. Future research directions

The findings of this review point toward several concrete directions for future research. First, greater emphasis is needed on hybrid human-AI assessment frameworks that balance automation efficiency with pedagogical accountability, particularly in summative contexts. Second, the development of task-specific and domain-adaptive LLMs may enhance robustness, fairness, and reliability in language-based assessment tasks. Third, multimodal AI approaches integrating textual, behavioral, and interaction data offer opportunities to capture richer representations of learning processes beyond academic performance alone, particularly in formative assessment contexts. Finally, strengthening reproducibility through shared datasets, standardized benchmarks, and transparent reporting practices is essential for advancing cumulative knowledge and responsible AI-based assessment research in higher education. In addition, future research would benefit from clearer and more consistent classification of AI-based assessments according to their primary purpose. Distinguishing whether AI-driven systems are intended to support formative learning processes, summative evaluation, or a combination of both would improve methodological transparency, enhance comparability across studies, and strengthen cumulative evidence in this field.

5. Conclusions and limitations

Advancements in AI, particularly ML, DL, and LLMs, have substantially influenced formative, summative, and both assessment practices in higher education disciplines. Across the reviewed studies, AI-driven techniques demonstrate strong potential for predicting academic outcomes, assessing performance and engagement, analyzing emotional and behavioral indicators, and supporting feedback and review processes. This SLR, covering studies published between 1997 and 2024, reveals a marked growth in AI-based educational assessment research since 2019, reflecting increasing institutional and scholarly interest in data-driven assessment practices. The predominant use of supervised and unsupervised techniques, most notably classification, regression, clustering, and generative models, highlights both the maturity of predictive analytics in education and the expanding role of language-based assessment tools. Analysis of publication and geographic trends further indicates that research activity is concentrated in Asia and North America, particularly in China and the United States, with computer science and engineering disciplines leading this work.

While AI-driven formative assessment has become well established as a tool for feedback, early intervention, and learning support, the comparatively limited and cautious use of AI in summative and other high-stakes assessment contexts reflects persistent concerns regarding reliability, transparency, fairness, and accountability. This imbalance underscores the need for balanced and hybrid assessment frameworks in which AI augments, rather than replaces, human judgment, particularly when assessment outcomes carry significant academic consequences.

Importantly, this review highlights that AI techniques in educational assessment function in complementary ways rather than as isolated solutions. Classification methods dominate due to their effectiveness in early warning systems and performance prediction, whereas regression and clustering approaches, though less frequently applied, offer valuable support for continuous outcome estimation and learner profiling. Generative AI and text mining approaches provide scalable, language-centered assessment and feedback capabilities, particularly in formative contexts; however, ethical considerations, privacy risks, and robustness limitations must be addressed before broader adoption in summative assessment settings. Together, these findings suggest that integrated, multi-technique assessment ecosystems may offer the most promising pathway for advancing precision, adaptability, and inclusiveness in higher education assessment.

From an institutional perspective, the findings emphasize the importance of investing in assessment infrastructures that support explainability, data governance, and instructor oversight. At the same time, the strong reliance on customized and institution-specific datasets raises concerns about generalizability and reproducibility, reinforcing the need to balance context-specific data with greater use of public, well-documented datasets and standardized reporting practices.

Despite its contributions, this review has several limitations that should be acknowledged. First, the scope is restricted to higher education, excluding insights from primary, secondary, and vocational education contexts. Second, only English-language publications were considered, which may limit geographic representation. Third, the review primarily centers on AI-supported prediction of student outcomes, rather than providing detailed comparative evaluations of assessment methods or algorithmic performance. Fourth, no systematic benchmarking of AI models was conducted, meaning conclusions regarding algorithm superiority remain tentative. Finally, although extensive database searches were conducted, limitations in keyword selection and database coverage may have resulted in the omission of relevant studies, and inconsistent reporting practices in primary studies occasionally constrained more detailed classification decisions of assessment purposes.

Overall, these limitations highlight important directions for future research, including broader educational contexts, more diverse linguistic and geographic coverage, comparative evaluation of AI models, and stronger open-science practices. Addressing these gaps will be essential for advancing robust, ethical, and pedagogically grounded AI-based assessment in higher education.

CRediT authorship contribution statement

Mohammed Bashraheel: Conceptualization, Methodology, Study design, Formal analysis, Data interpretation, Writing – original draft. **Gheorghita Ghinea:** Conceptualization, Methodology, Study design, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data for this article can be found online at doi:10.1016/j.cosrev.2026.100929.

Data availability

No data was used for the research described in the article.

References

[1] A. Abduljabbar, N. Gupta, L. Healy, Y. Kumar, J.J. Li, P. Morreale, A self-served AI tutor for growth mindset teaching, in: 2022 5th International Conference on Information and Computer Technologies (ICICT), IEEE, 2022, pp. 55–59.

[2] M. Adnan, M.I. Uddin, E. Khan, F.S. Alharithi, S. Amin, A.A. Alzahrani, Earliest possible global and local interpretation of students' performance in virtual learning environment by leveraging explainable AI, IEEE Access 10 (2022) 129843–129864.

[3] M. Afzaal, J. Nouri, A. Zia, P. Papapetrou, U. Fors, Y. Wu, et al., Automatic and intelligent recommendations to support students' self-regulation, in: 2021 International Conference on Advanced Learning Technologies (ICALT), IEEE, 2021, pp. 336–338.

[4] M. Afzaal, J. Nouri, A. Zia, P. Papapetrou, U. Fors, Y. Wu, et al., Explainable AI for data-driven feedback and intelligent action recommendations to support students self-regulation, Front. Artif. Intell. 4 (2021) 723447.

[5] M. Afzaal, A. Zia, J. Nouri, U. Fors, Informative feedback and explainable ai-based recommendations to support students' self-regulation, Technol. Knowl. Learn. 29 (1) (2024) 331–354.

[6] M. Ahamad, N. Ahmad, Students' knowledge assessment using the ensemble methods, Int. J. Inf. Technol. 13 (3) (2021) 1025–1032.

[7] K. Alalawi, R. Athauda, R. Chiong, An innovative framework to improve course and student outcomes, in: 2021 6th International Conference on Innovative Technology in Intelligent System and Industrial Applications (CITISIA), IEEE, 2021, pp. 1–6.

[8] B. Albreiki, N. Zaki, H. Alashwal, A systematic literature review of student' performance prediction using machine learning techniques, Educ. Sci. 11 (9) (2021) 552.

[9] S.I. Aleksandrovich, T. Ramazan, R. Utegaliyeva, B. Sarimbayeva, G. Keubassova, R. Bissalyyeva, et al., Transformative applications in Biology education: a case study on the efficacy of adaptive learning with numerical insights, Caspian J. Environ. Sci. 22 (2) (2024) 395–408.

[10] A. Ali, A. Deuter, L. Wehmeier, Personalized learning in automation: a 3d ai-based approach, in: 2023 IEEE Frontiers in Education Conference (FIE), IEEE, 2023, pp. 1–5.

[11] K. Ali, N. Barhom, F. Tamimi, M. Duggal, Chatgpt—a double-edged sword for healthcare education? Implications for assessments of dental students, Eur. J. Dent. Educ. 28 (1) (2024) 206–211.

[12] R. Alshabandar, A. Hussain, R. Keight, W. Khan, Students performance prediction in online courses using machine learning algorithms, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–7.

[13] B. AlShebli, S.A. Memon, J.A. Evans, T. Rahwan, China and the us produce more impactful AI research when collaborating together, Sci. Rep. 14 (1) (2024) 28576.

[14] F. Altoe, D. Joyner, Annotation-free automatic examination essay feedback generation. In 2019 IEEE learning with moocs (lwmoocs)(PP. 110–115), in: Proceedings of 2019 IEEE Learning with MOOCS, LWMOOCS 2019, 2019, pp. 110–115.

[15] R. Alubady, T.A. Diame, H. Sabah, H.H.J. Mahdi, M. Saleem, K. Cengiz, et al., Anticipating student engagement in classroom through iot-enabled intelligent teaching model enhanced by machine learning, Fusion: Pract. Appl. 13 (1) (2023) 189–202.

[16] K. Ananiadoui, M. Claro, 21st Century Skills and Competences for New Millennium Learners in Oecd Countries, 2009.

[17] A.W. Astin, Assessment for Excellence: the Philosophy and Practice of Assessment and Evaluation in Higher Education, Rowman & Littlefield Publishers, 2012.

[18] I. Atmosukarto, C.W. Sin, P. Iyer, N.H. Tong, K.W.P. Yu, Enhancing adaptive online Chemistry course with ai-chatbot, in: 2021 IEEE International Conference on Engineering, Technology & Education (TALE), IEEE, 2021, pp. 838–843.

[19] Z.Z. Aung, G.R. Shinha, Unveiling insights: a machine learning approach to decipher student sentiments in computer university of Myanmar, in: 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE), IEEE, 2024, pp. 1–5.

[20] D. Babichenko, M. Druzdzel, N. Benedict, G. Tabas, J.B. McGee, Moving beyond branching: evaluating educational impact of procedurally-generated virtual patients, in: 2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH), IEEE, 2019, pp. 1–8.

[21] M. Bahrehvar, M. Moshirpour, Agile teaching: automated student support and feedback generation, in: 2023 IEEE Frontiers in Education Conference (FIE), IEEE, 2023, pp. 1–9.

[22] R.S. Baker, K. Yacef, The state of educational data mining in 2009: a review and future visions, J. Educ. Data Min. 1 (1) (2009) 3–17.

[23] M.TP. Beerepoot, Formative and summative automated assessment with multiple-choice question banks, J. Chem. Educ. 100 (8) (2023) 2947–2955.

[24] B. Bell, B. Cowie, The characteristics of formative assessment in science education, Sci. Educ. 85 (5) (2001) 536–553.

[25] B.S. Bloom, et al., Handbook on Formative and Summative Evaluation of Student Learning, 1971.

[26] P. Broadfoot, P. Black, Redefining assessment? The first ten years of assessment in education, Assess. Educ.: Princ. Policy Pract. 11 (1) (2004) 7–26.

[27] J. Broisin, C. Hérouard, Design and evaluation of a semantic indicator for automatically supporting programming learning, in: EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining 2019.

[28] H.D. Brown, P. Abeywickrama, Language Assessment: Principles and Classroom Practices, Pearson, 2019.

[29] M.-Y. Cai, J.-Y. Wang, G.-D. Chen, J.-H. Wang, S.-H. Yang, A digital reality theater with the mechanisms of real-time spoken language evaluation and interactive switching of scenario & virtual costumes: effects on motivation and learning performance, in: 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT), IEEE, 2020, pp. 295–299.

[30] D.R. Carless, On your marks: Learner-focused feedback practices and feedback literacy (2020).

[31] C. Chaiprasurt, R. Amornchewin, P. Kunpitak, Using motivation to improve learning achievement with a chatbot in blended learning, World J. Educ. Technol. Curr. 14 (4) (2022) 1133–1151.

[32] C.-Y. Chang, S.-Y. Kuo, G.-H. Hwang, Chatbot-facilitated nursing education, Educ. Technol. Soc. 25 (1) (2022) 15–27.

[33] S. Chida, K. Minamino, Effectiveness of relative evaluation feedback using motivation tests and learning data in university Mathematics, in: 2023 11th International Conference on Information and Education Technology (ICIET), IEEE, 2023, pp. 405–410.

[34] T.B. Chindhe, B.M. Patil, Student performance appraisal system, in: 2021 6th International Conference on Communication and Electronics Systems (ICCES), IEEE, 2021, pp. 930–935.

[35] K. Cochran, C. Cohn, J.F. Rouet, P. Hastings, Improving automated evaluation of student text responses using gpt-3.5 for text data augmentation, in: International Conference on Artificial Intelligence in Education, vol. 13916 LNAI, Springer, 2023, pp. 217–228.

[36] M. Cogliano, M.L. Bernacki, J.C. Hilpert, C.L. Strong, A self-regulated learning analytics prediction-and-intervention design: detecting and supporting struggling Biology students, J. Educ. Psychol. 114 (8) (2022) 1801–1816.

[37] G.T. Crisp, Integrative assessment: reframing assessment practice for current and future learning, Assess. Eval. High. Educ. 37 (1) (2012) 33–43.

[38] N. Dehbozorgi, M.T. Kunuku, An llm-based reflection analysis tool for identifying and addressing challenging topics, in: Proceedings of the 55th ACM Technical Symposium on Computer Science Education v. 2, vol. 2, 2024, pp. 1618–1619.

[39] N. Dehbozorgi, S. MacNeil, Semi-automated analysis of reflections as a continuous course, in: 2019 IEEE Frontiers in Education Conference (FIE), volume 2019-October, IEEE, 2019, pp. 1–5.

[40] N. Dehbozorgi, M.L. Maher, M. Dorodchi, Sentiment analysis on conversations in collaborative active learning as an early predictor of performance, in: 2020 IEEE Frontiers in Education Conference (FIE), volume 2020-October, IEEE, 2020, pp. 1–9.

[41] K. Delgado, J.M. Origgi, T. Hasanpoor, H. Yu, D. Allessio, I. Arroyo, et al., Student engagement dataset, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, volume 2021-October, 2021, pp. 3621–3629.

[42] S. Disa, A.M. Idkhan, Web e-learning: automated essay assessment based on natural language processing using vector space model, in: 2022 4th International Conference on Cybernetics and Intelligent System (ICORIS), IEEE, 2022, pp. 1–4.

[43] J. Dolin, P. Black, W. Harlen, A. Tiberghien, Exploring relations between formative and summative assessment, Contrib. Sci. Educ. Res. 4 (2018) 53–80.

[44] L.M. Díaz, C.Z. Chevez, Recounting the history of user interface research through publication titles, in: Proceedings of the Latin American Conference on Human Computer Interaction, 2015, pp. 1–4.

[45] D.M. Elbourhamy, Automated sentiment analysis of visually impaired students' audio feedback in virtual learning environments, PeerJ Comput. Sci. 10 (2024) e2143.

[46] M. Ergezer, B. Kucharski, A. Carpenter, Work in progress: designing laboratory work for a novel embedded AI course, in: 2018 ASEE Annual Conference & Exposition, Volume 2018-June, ASEE Conferences, 2018.

[47] J. Escalante, A. Pack, A. Barrett, Ai-generated feedback on writing: insights into efficacy and enl student preference, Int. J. Educ. Technol. High. Educ. 20 (1) (2023) 57.

[48] F. Fang, X. Jiang, The analysis of artificial intelligence digital technology in art education under the internet of things, IEEE Access 12 (2024) 22928–22937.

[49] A.M. Fazlollahi, R. Yilmaz, A. Winkler-Schwartz, N. Mirchi, N. Ledwos, M. Bakhaidar, et al., AI in surgical curriculum design and unintended outcomes for technical competencies in simulation training, JAMA Network Open 6 (9) (2023) e2334658.

[50] A. Fink, Conducting Research Literature Reviews: from the Internet to Paper, Sage Publications, 2019.

[51] R.B. Fletcher, L.H. Meyer, H. Anderson, P. Johnston, M. Rees, Faculty and students conceptions of assessment in higher education, Higher Educ. 64 (1) (2012) 119–133.

[52] T.S. Francišković, A. Anđelić, J. Slivka, N. Luburić, A. Kovačević, Predicting students' final exam scores based on their regularity of engagement with pre-class activities in a flipped classroom, Int. Conf. Comput. Support. Educ, CSEDU - Proc. 2 (2024) 97–107.

[53] V. Franzoni, A. Milani, P. Mengoni, F. Piccinato, Artificial intelligence visual metaphors in e-learning interfaces for learning analytics, Appl. Sci. 10 (20) (2020) 7195.

[54] R. Gao, H.E. Merzdorf, S. Anwar, M.C. Hipwell, A.R. Srinivasa, Automatic assessment of text-based responses in post-secondary education: a systematic review, Comput. Educ.: Artif. Intell. 6 (2024) 100206.

[55] R. Gessner, J. Hargis, J. Wade, Slidespace: a social realtime teaching and learning environment, in: 2021 IEEE Frontiers in Education Conference (FIE), volume 2021-October, IEEE, 2021, pp. 1–4.

[56] R. Glaser, N. Chudowsky, J.W. Pellegrino, Knowing what Students Know: the Science and Design of Educational Assessment, National Academies Press, 2001.

[57] N. Glazer, Formative plus summative assessment in large undergraduate courses: why both? Int. J. Teach. Learn. High. Educ. 26 (2) (2014) 276–286.

[58] M.J. Grant, A. Booth, A typology of reviews: an analysis of 14 review types and associated methodologies, Health Inf. Libr. J. 26 (2) (2009) 91–108.

[59] F. Grivokostopoulou, I. Perikos, I. Hatzilygeroudis, An educational system for learning search algorithms and automatically assessing student performance, Int. J. Artif. Intell. Educ. 27 (1) (2017) 207–240.

[60] S. Gupta, R.R. Dharamshi, V. Kakde, An impactful and revolutionized educational ecosystem using generative AI to assist and assess the teaching and learning benefits, fostering the post-pandemic requirements, in: 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE), IEEE, 2024, pp. 1–4.

[61] S.K. Gupta, T.S. Ashwin, R.M.R. Guddeti, Students' affective content analysis in smart classroom environment using deep learning techniques, Multimed. Tools Appl. 78 (18) (2019) 25321–25348.

[62] W. Harlen, Teachers' summative practices and assessment for learning – tensions and synergies, Curric. J. 16 (2) (2005) 207–223.

[63] V. Hegde, N. Surendran, M. Vaishnavi, Predicting student failure using peer-based evaluation and ratings, in: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, 2023, pp. 1–6.

[64] S. Hellman, M. Rosenstein, A. Gorman, W. Murray, L. Becker, A. Baikadi, et al., Scaling up writing in the curriculum: batch mode active learning for automated essay scoring, in: Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale, 2019, pp. 1–10.

[65] J. Hernandez-Gonzalez, P.J. Herrera, On the supervision of peer assessment tasks: an efficient instructor guidance technique, IEEE Trans. Learn. Technol. 16 (6) (2023) 926–939.

[66] W. Ho, D. Lee, Enhancing engineering education in the roblox metaverse: utilizing Chatgpt for game development for electrical machine course, Int. J. Adv. Sci. Eng. Inf. Technol. 13 (3) (2023) 1052–1058.

[67] Q.N. Hong, S. Fàbregues, G. Bartlett, F. Boardman, M. Cargo, P. Dagenais, et al., The mixed methods appraisal tool (mmat) version 2018 for information professionals and researchers, Educ. Inf. 34 (4) (2018) 285–291.

[68] Y.-H. Hu, J.S. Fu, H.-C. Yeh, Developing an early-warning system through robotic process automation: are intelligent tutoring robots as effective as human teachers? Interact. Learn. Environ. 32 (6) (2023) 2803–2816.

[69] H. David, I.V. Smith, C. Zilles, Code Generation Based Grading: Evaluating an Auto-Grading Mechanism for "Explain-in-Plain-English" Questions, vol. 1, Association for Computing Machinery, 2024, pp. 171–177.

[70] Q. Jia, M. Young, Y. Xiao, J. Cui, C. Liu, P. Rashid, et al., Insta-reviewer: a data-driven approach for generating instant feedback on students' project reports, Int. Educ. Data Min. Soc., 2022, https://doi.org/10.5281/zenodo.6853099

[71] D. Joyner, Intelligent evaluation and feedback in support of a credit-bearing MOOC, in: Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part II 19, vol. 10948 LNAI, Springer, 2018, pp. 166–170.

[72] S.L. Kanuru, M. Priyaadharshini, Lifelong learning in higher education using learning analytics, Proc. Comput. Sci. 172 (2020) 848–852.

[73] P.S. Kasliwal, R. Gunjan, V. Shete, Contextual emotion detection of e-learners for recommendation system, J. Eng. Educ. Transform. 37 (2) (2023) 200–208.

[74] S. Keele, Guidelines for performing systematic literature reviews in software engineering, EBSE Technical Report EBSE-2007-01, School of Computer Science and Mathematics, Keele University, UK, 2007.

[75] I. Koprinska, J. Stretton, K. Yacef, Predicting student performance from multiple data sources, in: Artificial Intelligence in Education: 17th International Conference, AIED 2015, Madrid, Spain, June 22-26, 2015. Proceedings 17, vol. 9112, Springer, 2015, pp. 678–681.

[76] S. Kushnarev, K. Kang, S. Goyal, Assessing the efficacy of personalized online homework in a first-year engineering multivariate calculus course, in: 2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), IEEE, 2020, pp. 770–773.

[77] P. Lai, J. Chen, V. Man, C.H. Chan, A new frontier in ai-assisted English oral presentation assessment, in: 2023 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE), IEEE, 2023, pp. 1–8.

[78] A.V.Y. Lee, Supporting students' generation of feedback in large-scale online course with artificial intelligence-enabled evaluation, Stud. Educ. Eval. 77 (2023) 101250.

[79] A.V.Y. Lee, A.C. Luco, S.C. Tan, A human-centric automated essay scoring and feedback system for the development of ethical reasoning, Educ. Technol. Soc. 26 (1) (2023) 147–159.

[80] H.-Y. Lee, P.-H. Chen, W.-S. Wang, Y.-M. Huang, T.-T. Wu, Empowering Chatgpt with guidance mechanism in blended learning: effect of self-regulated learning, higher-order thinking skills, and knowledge construction, Int. J. Educ. Technol. High. Educ. 21 (1) (2024) 16.

[81] Y. Li, The digital transformation of college English classroom: application of artificial intelligence and data science, EAI Endorsed Trans. on Scalable Inf. Syst. 11 (5) (2024).

[82] Y. Li, H. Liu, X. Bai, Q. Li, M. Cai, J. Wang, The impact of classroom learning behavior on learning outcomes: a computer vision study, in: Proceedings of the 9th International Conference on Education and Training Technologies, 2023, pp. 1–8.

[83] J. Lin, W. Dai, L.-A. Lim, Y.-S. Tsai, R.F. Mello, H. Khosravi, et al., Learner-centred analytics of feedback content in higher education, in: LAK23: 13th International Learning Analytics and Knowledge Conference, 2023, pp. 100–110.

[84] J. Liu, Enhancing English language education through big data analytics and generative AI, J. Web Eng. 23 (2) (2024) 227–250.

[85] Z. Liu, W. Qian, L. Wang, Evaluation of the impact of artificial intelligence on college students' learning, in: 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), IEEE, 2023, pp. 1–8.

[86] A. Magooda, D. Litman, A. Ashraf, M. Menekse, Improving the quality of students' written reflections using natural language processing: model design and classroom evaluation, in: International Conference on Artificial Intelligence in Education, vol. 13355 LNCS, Springer, 2022, pp. 519–525.

[87] W. Mahfouz, H.-D. Wuttke, Deep-learning api-feature specification tool for formative assessment in workshops, in: 2022 IEEE Frontiers in Education Conference (FIE), volume 2022-October, IEEE, 2022, pp. 1–9.

[88] D. Maia, S.C. dos Santos, Monitoring students' professional competencies in PBL: a proposal founded on constructive alignment and supported by AI technologies, in: 2022 IEEE Frontiers in Education Conference (FIE), vol. 2022-October, IEEE, 2022, pp. 1–8.

[89] C. Maimone, B.M. Dolan, M.M. Green, S.M. Sanguino, P.M. Garcia, C.L. O'brien, Utilizing natural language processing of narrative feedback to develop a predictive model of pre-clerkship performance: lessons learned, Perspect. Med. Educ. 12 (1) (2023) 141–148.

[90] R.S. Malaquias, I.M.B. Filho, Middleware for healthcare systems: a systematic mapping, Lect. Notes Comput. Sci. 12957 LNCS (2021) 394–409.

[91] M.M. Mariani, N. Hashemi, J. Wirtz, Artificial intelligence empowered conversational agents: a systematic literature review and research agenda, J. Bus. Res. 161 (2023).

[92] D.M. Mertens, Research and Evaluation in Education and Psychology: Integrating Diversity with Quantitative, Qualitative, and Mixed Methods, Sage Publications, 2023.

[93] D. Moher, L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, et al., Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement, Syst. Rev. 4 (1) (2015) 1–9.

[94] M. Morales-Chan, H.R. Amado-Salvatierra, J.A. Medina, R. Barchino, R. Hernández-Rizzardini, A.M. Teixeira, Personalized feedback in massive open online courses: harnessing the power of langchain and openai API, Electronics 13 (10) (2024) 1960.

[95] R. Morris, T. Perry, L. Wardle, Formative assessment and feedback for learning in higher education: a systematic review, Rev. Educ. 9 (3) (2021) e3292.

[96] Z. Munn, M.D.J. Peters, C. Stern, C. Tufanaru, A. McArthur, E. Aromataris, What kind of systematic review should i conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences, BMC Med. Res. Methodol. 18 (1) (2018) 1–9.

[97] R. Murali, N. Ravi, A. Surendran, Augmenting virtual labs with artificial intelligence for hybrid learning, in: 2024 IEEE Global Engineering Education Conference (EDUCON), IEEE, 2024, pp. 1–10.

[98] G. Nalli, R. Culmone, A. Perali, D. Amendola, Online tutoring system for programming courses to improve exam pass rate, J. e-Learn. Knowl. Soc. 19 (1) (2023) 27–35.

[99] S. Nayak, R. Agarwal, S.K. Khatri, M. Mohammadian, Student outcome assessment on structured query language using rubrics and automated feedback generation, Int. J. Adv. Comput. Sci. Appl. 15 (3) (2024) 728–736.

[100] S. Nicoll, K. Douglas, C. Brinton, Giving feedback on feedback: an assessment of grader feedback construction on student performance, in: LAK22: 12th International Learning Analytics and Knowledge Conference, 2022, pp. 239–249.

[101] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, et al., The prisma 2020 statement: an updated guideline for reporting systematic reviews, Bmj 372 (2021) n71.

[102] M.J. Page, D. Moher, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, et al., Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews, bmj 372 (2021) n160.

[103] S. Pahlevan-Sharif, P. Mura, S.N.R. Wijesinghe, A systematic review of systematic reviews in tourism, J. Hosp. Tour. Manag. 39 (2019) 158–165.

[104] G. Paravati, F. Lamberti, V. Gatteschi, C. Demartini, P. Montuschi, Point cloud-based automatic assessment of 3d computer animation courseworks, IEEE Trans. Learn. Technol. 10 (4) (2016) 532–543.

[105] R. Parikh, H. Nimonkar, R. Gandhi, T. Budhrani, A. Dalvi, I. Siddavatam, Streamlining educational assessment: a user-centric analysis of an ai-powered examination app, in: 2023 6th International Conference on Advances in Science and Technology (ICAST), IEEE, 2023, pp. 341–345.

[106] L. Patel, A. Alsobeh, Slangllm: dynamic detection and contextual filtering of slang in NLP applications, in: Proceedings of the 1st International Conference on Secure IoT, Assured and Trusted Computing (SATC), IEEE, 2025, pp. 1–6.

[107] D.F. Polit, Getting serious about test-retest reliability: a critique of retest research and some recommendations, Qual. Life Res. 23 (6) (2014) 1713–1720.

[108] F. Pramudianto, T. Chhabra, E.F. Gehringer, C. Maynards, Assessing the quality of automatic summarization for peer review in education, in: EDM (Workshops), vol. 1633, Citeseer, 2016, pp. 1–5.

[109] R. Pranckutė, Web of science (wos) and scopus: the titans of bibliographic information in today's academic world, Publications 9 (1) (2021).

[110] C. Qiao, X. Hu, Leveraging semantic facets for automatic assessment of short free text answers, IEEE Trans. Learn. Technol. 16 (1) (2022) 26–39.

[111] R. Raman, P. Singh, V.K. Singh, R. Vinuesa, P. Nedungadi, Understanding the bibliometric patterns of publications in IEEE access, IEEE Access 10 (2022) 35561–35577.

[112] V. Ramasamy, D. Singsen, G.S. Walia, Fostering student engagement and success in STEM education: an ai-driven exploration of high impact practices from cross-disciplinary general education courses, J. Eng. Educ. Transform. 37 (Special Issue 2) (2024) 849–857.

[113] M.P. Rashid, E. Gehringer, H. Khosravi, Navigating (dis) agreement: AI assistance to uncover peer feedback discrepancies, in: Proceedings of the 14th Learning Analytics and Knowledge Conference, 2024, pp. 907–914.

[114] A.A. Rekhi, Metamorphosing teaching and learning with python: a practical guide for teachers employing text data analysis using the bag of words (BOW) model, in: 2023 International Conference on Quantum Technologies, Communications, Computing, Hardware and Embedded Systems Security (iQ-CCHESS), IEEE, 2023, pp. 1–8.

[115] J. Renzella, A. Cain, J.-G. Schneider, Verifying student identity in oral assessments with deep speaker, Comput. Educ.: Artif. Intell. 3 (2022) 100044.

[116] J. Rodriguez-Ruiz, A. Alvarez-Delgado, P. Caratozzolo, Use of natural language processing (NLP) tools to assess digital literacy skills, in: 2021 Machine Learning-Driven Digital Technologies for Educational Innovation Workshop, IEEE, 2021, pp. 1–8.

[117] C. Romero, S. Ventura, Educational data mining: a survey from 1995 to 2005, Expert Syst. Appl. 33 (1) (2007) 135–146.

[118] O.L. Dos Santos, D. Cury, Challenging the confirmation bias: using Chatgpt as a virtual peer for peer instruction in computer programming education, in: 2023 IEEE Frontiers in Education Conference (FIE), IEEE, 2023, pp. 1–7.

[119] M. Selmi, H. Hage, E. Aïmeur, Evaluating lsa sensibility to disclosure in learners' interactions, in: 2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA), IEEE, 2015, pp. 1–8.

[120] R. Sembey, R. Hoda, J. Grundy, Emerging technologies in higher education assessment and feedback practices: a systematic literature review, J. Syst. Softw. 211 (2024) 111988.

[121] K. Sharma, Z. Papamitsiou, M. Giannakos, Building pipelines for educational data using AI and multimodal analytics: a "grey-box" approach, Br. J. Educ. Technol. 50 (6) (2019) 3004–3031.

[122] S.J. Shi, J.W. Li, R. Zhang, A study on the impact of generative artificial intelligence supported situational interactive teaching on students''flow'experience and learning effectiveness—a case study of legal education in China, Asia Pac. J. Educ. 44 (1) (2024) 112–138.

[123] M. Songxia, J. Xiaoping, Evaluation of overseas students' performance in Chinese courses using statistical learning, in: 2011 International Conference on E-Business and E-Government (ICEE), IEEE, 2011, pp. 1–4.

[124] S.E. Sorour, T. Mine, K. Godaz, S. Hirokawax, Comments data mining for evaluating student's performance, in: 2014 IIAI 3rd International Conference on Advanced Applied Informatics, IEEE, 2014, pp. 25–30.

[125] M. Staples, M. Niazi, Experiences using systematic review guidelines, J. Syst. Softw. 80 (9) (2007) 1425–1437.

[126] P. Stutz, M. Elixhauser, J. Grubinger-Preiner, V. Linner, E. Reibersdorfer-Adelsberger, C. Traun, et al., CH (e) atgpt? An anecdotal approach addressing the impact of Chatgpt on teaching and learning giscience, GI Forum 11 (1) (2023) 140–147.

[127] Z. Sun, Q. Hu, R. Gupta, R. Zemel, Y. Xu, Toward informal language processing: knowledge of slang in large language models, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Association for Computational Linguistics, 2024, pp. 1683–1701.

[128] Z. Swiecki, A.R. Ruis, D. Gautam, V. Rus, D.W. Shaffer, Understanding when students are active-in-thinking through modeling-in-context, Br. J. Educ. Technol. 50 (5) (2019) 2346–2364.

[129] M.C. Sáiz-Manzanares, R. Marticorena-Sánchez, L.J. Martín-Antón, I.G. Díez, L. Almeida, Perceived satisfaction of university students with the use of chatbots as a tool for self-regulated learning, Heliyon 9 (1) (2023).

[130] M. Taras, Using assessment for learning and learning from assessment, Assess. Eval. High. Educ. 27 (6) (2002) 501–510.

[131] M. Taras, Summative assessment: the missing link for formative assessment, J. Furth. High. Educ. 33 (1) (2009) 57–69.

[132] C.C. Tossell, N.L. Tenhundfeld, A. Momen, K. Cooley, E.J. de Visser, Student perceptions of Chatgpt use in a college essay assignment: implications for learning, grading, and trust in artificial intelligence, IEEE Trans. Learn. Technol. 17 (2024) 1069–1081.

[133] W. Villegas-Ch, J. García-Ortiz, I. Urbina-Camacho, A. Mera-Navarrete, Proposal for a system for the identification of the concentration of students who attend online educational models, Computers 12 (4) (2023) 74.

[134] W. Villegas-Ch, J. Govea, R. Gurierrez, A. Mera-Navarrete, Improving interaction and assessment in hybrid educational environments: an integrated approach in Microsoft teams with the use of AI techniques, IEEE Access 12 (2024) 93723–93738.

[135] C.S. Wagner, L. Leydesdorff, Network structure, self-organization, and the growth of international collaboration in science, Res. Policy 34 (10) (2005) 1608–1618.

[136] K. Wang, E.Y. Fu, G. Ngai, H.V. Leong, Identifying key learning factors in service-leaning programs using machine learning, in: 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC), IEEE, 2022, pp. 1312–1317.

[137] L. Wang, X. Chen, C. Wang, L. Xu, R. Shadiev, Y. Li, Chatgpt's capabilities in providing feedback on undergraduate students' argumentation: a case study, Think. Skills Creat. 51 (2024) 101440.

[138] Z. Wang, Adaptability of the reform of speaking teaching mode of master's foreign language based on virtual simulation technology, J. Electr. Syst. 20 (2024) 551–561.

[139] C. Wangwiwattana, Y. Tongvivat, Semi-automatic short-answer grading tools for Thai language using natural language processing, in: Proceedings of the 2022 5th International Conference on Education Technology Management, 2022, pp. 123–128.

[140] C. Wangwiwattana, Y. Tongvivat, Automating academic assessment: a large language model approach, in: 2023 7th International Conference on Information Technology (InCIT), IEEE, 2023, pp. 330–334.

[141] H. Wei, Z. Li, H. Xie, K. Hung, M. Wang, Beyond scores: a novel method for predicting student performance based on rank and positional embedding, in: 2023 10th International Conference on Behavioural and Social Computing (BESC), IEEE, 2023, pp. 1–7.

[142] D. Wiliam, P. Black, Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? Br. Educ. Res. J. 22 (5) (1996) 537–548.

[143] N. Winstone, D. Carless, Designing Effective Feedback Processes in Higher Education: A Learning-Focused Approach, Routledge, 2019.

[144] J. Wong, O. Viberg, Supporting self-regulated learning with generative AI: a case of two empirical studies, in: LAK Workshops, vol. 3667, 2024, pp. 223–229.

[145] Y. Xiao, G. Zingle, Q. Jia, S. Akbar, Y. Song, M. Dong, et al., Problem detection in peer assessments between subjects by effective transfer learning and active learning, Int. Educ. Data Min. Soc. (2020) 516–523.

[146] M. Yorke, Formative assessment in higher education: moves towards theory and the enhancement of pedagogic practice, Higher Educ. 45 (4) (2003) 477–501.

[147] T. Yuan, Z. Wang, P.-L.P. Rau, Design of intelligent real-time feedback system in online classroom, in: International Conference on Human-Computer Interaction, vol. 14024 LNCS, Springer, 2023, pp. 326–335.

[148] O. Zawacki-Richter, V.I. Marín, M. Bond, F. Gouverneur, Systematic review of research on artificial intelligence applications in higher education–where are the educators? Int. J. Educ. Technol. High. Educ. 16 (1) (2019) 1–27.

[149] X. Zhai, Y. Yin, J.W. Pellegrino, K.C. Haudek, L. Shi, Applying machine learning in science assessment: a systematic review, Stud. Sci. Educ. 56 (1) (2020) 111–151.

[150] P. Zhang, G. Tur, A systematic review of Chatgpt use in k–12 education, Eur. J. Educ. 59 (2) (2024) e12599.

[151] X. Zhang, J. Sun, Y. Deng, Design and application of intelligent classroom for English language and literature based on artificial intelligence technology, Appl. Artif. Intell. 37 (1) (2023) 2216051.

[152] Y. Zhang, E.F. Gehringer, Can students produce effective training data to improve formative feedback? in: 2021 IEEE Frontiers in Education Conference (FIE), vol. 2021-October, IEEE, 2021, pp. 1–7.

[153] Z. Zhao, A new cloud computing-based assessment of issues in online teaching management in the post-epidemic era of Covid-19, Int. J. Recent Innov. Trends Comput. Commun. 11 (6 S) (2023) 138–151.

[154] C. Zheng, X. Chen, H. Zhang, C.S. Chai, Automated versus peer assessment: effects of learners' English public speaking, Language Learning and Technology 28 (2) (2024) 210–228.

[155] Y. Zhou, Z. Zeng, H. Wang, Using spectral clustering association algorithm upon teaching big data for precise education, Math. Probl. Eng. 2022 (1) (2022) 7214659.

[156] Y. Zhuang, L. Wang, M. Zhang, S. Lin, H. Hu, X. Tao, Optes: a tool for behavior-based student programming progress estimation, in: 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), vol. 2023-June, IEEE, 2023, pp. 122–131.

[157] C. Şerban, L. Ioan, Qlearn: towards a framework for smart learning environments, Proc. Comput. Sci. 176 (2020) 2812–2821.