

Measurement Innovations in Artificial Intelligence Based Educational Assessment: A Hybrid Systematic Literature Review and Bibliometric Analysis

Rustam AR. Selang¹, Awaluddin Tjalla¹, Iva Sarifah¹

¹State University of Jakarta, Indonesia

rus.selang85@gmail.com; awaluddin.tjalla@gmail.com; ivasarifah@unj.ac.id

Proceedings of the Kalijaga International Conference on Education 2026

Abstract. Artificial Intelligence-based educational assessment has increasingly become an important area of educational research, particularly in relation to measurement innovation. The development of AI in assessment has expanded beyond automation and efficiency, raising critical issues concerning validity, reliability, fairness, interpretability, human oversight, and alignment with human judgment. This study aims to map and synthesize measurement innovations in Artificial Intelligence-based educational assessment through a hybrid approach that combines a systematic literature review and bibliometric analysis. Specifically, this study identifies research trends, dominant themes, forms of innovation, methodological issues, and future research agendas in AI-based educational measurement. This study used scholarly articles from the Scopus database published between 2021 and 2026. The article selection process followed the stages of identification, screening, eligibility assessment, and inclusion, resulting in 107 relevant articles. Bibliometric analysis was employed to map publication trends, dominant keywords, sources, authors, institutions, countries, and patterns of scholarly collaboration. Meanwhile, the systematic literature review was used to synthesize key themes, innovation models, methodological concerns, and future directions. The findings show that research on AI-based educational assessment developed rapidly during the 2023–2026 period. The major themes identified include automated scoring, automated feedback, multimodal assessment, learning analytics, and human-aligned validation. The literature also indicates that AI-based measurement innovations are not limited to improving assessment efficiency but also emphasize the quality and trustworthiness of assessment processes, particularly in relation to validity, reliability, fairness, interpretability, human oversight, and consistency with human assessment. This study concludes that the development of AI-based educational assessment is shifting from assessment automation toward trustworthy AI-based educational measurement. Such measurement should be valid, transparent, multimodal, and supported by meaningful human involvement. The findings provide a conceptual and bibliometric foundation for researchers, educators, and policymakers in developing AI-based educational assessment systems that are methodologically robust, ethically responsible, and aligned with educational measurement principles.

Keywords: *Educational Assessment, Artificial Intelligence, Measurement Innovation, Systematic Literature Review, Bibliometric Analysis.*

INTRODUCTION

The development of Artificial Intelligence (AI) has transformed educational assessment, especially how measurement is conducted, validated, and interpreted for academic decision-making. Assessment practices once grounded in manual grading, conventional rubrics, and outcome-oriented measurement have shifted toward automated, adaptive, multimodal, and data-driven systems. Prior studies show that AI can support item evaluation and generation, distractor development, examination evaluation, cognitive-level classification, automated scoring, writing evaluation, and discourse coherence assessment [1], [2], [3], [4], [5], [6], [7], [8]. These developments indicate that AI is no longer merely a technical aid but part of a complex educational measurement system. Nevertheless, AI-based measurement depends on construct validity, rubric quality, data consistency, score interpretability, and alignment with human judgment. Therefore, such innovation must be examined for efficiency, validity, reliability, fairness, and accountability.

This study is motivated by a gap between AI's technical capabilities and the suitability of its measurement outcomes for educational decisions. Although GPT-based systems may correlate highly with human scoring, stability and consistency problems remain [9]. Automated metrics may fail to capture contextual quality, including culture, ideology, fluency, and coherence, and may not align with human

evaluation [10], [11], [12]. Automated indicators may also encourage metric gaming when they become optimization targets [13]. These findings confirm that automation does not necessarily produce valid measurement. In education, this issue is critical because scores are used for diagnosis, feedback, ability mapping, and decisions. Thus, AI-based measurement research must address implementation models, validity, reliability, interpretability, and human oversight.

Recent literature shows that AI-based assessment is expanding into multimodal assessment, learning analytics, automated feedback, performance evaluation, and human-aligned validation. Studies have examined image-text measurement, multimodal diagram evaluation, creativity scoring, procedural skill micro-assessment, and domain-specific automated evaluation [14], [15], [16]. In digital learning, AI has supported analytics, online discussions, project-based learning, gamified learning, peer assessment, and calibration [17], [18], [19], [20], [21]. Yet these developments face limitations involving generalizability, bias, data quality, and pedagogical interpretation. Concerns about LLM-as-judge bias, faithfulness, relevance, safety, usefulness, and inauthentic text detection reinforce human-in-the-loop governance [22], [23].

Previous research remains fragmented across domains, methods, objects, and AI technologies. Studies vary from item quality, writing assessment, and open-ended responses to metric validation, human judgment, accessibility, representativeness, evaluation culture, and algorithmic evaluation [24], [4], [25], [26], [27], [28], [29], [30], [31], [32], [33]. This fragmentation shows that AI-based measurement is technical, institutional, ethical, and methodological. A hybrid systematic literature review and bibliometric analysis is therefore needed to integrate quantitative mapping with qualitative synthesis.

This study maps and synthesizes AI-based educational measurement innovation through systematic review and bibliometric analysis. It examines publication trends, collaboration patterns, themes, conceptual clusters, knowledge evolution, innovation models, psychometric quality, human oversight, gaps, novelty, and future agendas. This focus reflects the shift from automated scoring toward trustworthy AI-based educational measurement. The study contributes conceptually by clarifying AI-based assessment, methodologically by combining bibliometric mapping with systematic synthesis, and practically by identifying principles for valid, transparent, fair, accountable, and human-aligned systems. Overall, AI is positioned not as a replacement for human judgment but as a measurement support system governed through rubrics, validation, interpretability, and governance.

METHODS

This study employed a hybrid review design, combining a Systematic Literature Review (SLR) and bibliometric analysis to map, synthesize, and evaluate the literature on measurement innovation in Artificial Intelligence-based educational assessment. The SLR design was used to identify the forms, functions, models, and issues related to validity, reliability, fairness, interpretability, human oversight, and the alignment of AI with human assessment. Meanwhile, bibliometric analysis was used to map publication trends, journal sources, authors, institutions, countries, keywords, thematic clusters, and patterns of scientific collaboration.

The research sample consisted of scholarly articles obtained from the Scopus database through a search conducted on May 8, 2026. The search strategy was applied to titles, abstracts, and keywords using the following search string: (“AI Assessment” AND “Educational Measurement”) OR (“Automated Evaluation”). The selection process is presented in the PRISMA flow diagram below.

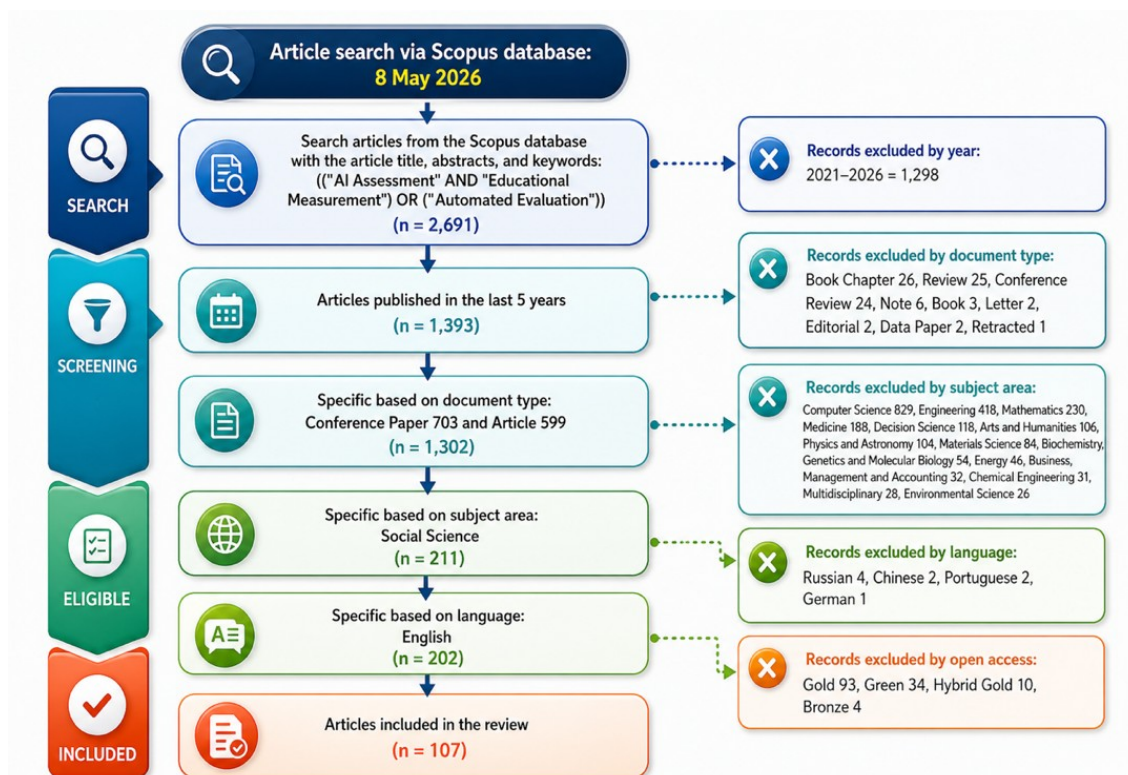


Figure 1. Systimatic Literature Review Information Flow Using PRISMA

Based on the PRISMA flow diagram above, the initial search yielded 2,691 documents. The documents were then filtered based on the publication period of 2021–2026, resulting in 1,393 articles. The next stage limited the document types to conference papers and articles, producing 1,302 documents. Further screening based on the Social Sciences subject area resulted in 211 documents. After applying the English-language criterion, 202 documents remained. The final selection resulted in 107 articles included in the review.

The research instruments consisted of an article selection protocol, a PRISMA flow diagram, a data extraction sheet, a list of inclusion and exclusion criteria, and a literature synthesis matrix. The inclusion criteria covered articles that discussed AI, automated evaluation, automated scoring, educational measurement, assessment innovation, learning analytics, multimodal assessment, or human-aligned validation in the context of educational assessment. The exclusion criteria included documents outside the specified year range, documents that were not articles or conference papers, documents outside the Social Sciences subject area, non-English documents, and documents that were not relevant to the focus of assessment measurement.

The data collection procedure was conducted in four stages: identifying articles from Scopus, screening based on year and document type, assessing eligibility based on subject area and language, and finalizing the articles that met the criteria for analysis. Each selected article was extracted based on author name, year, title, research focus, theory or framework used, method, main findings, limitations, and relevance to the study.

Bibliometric analysis was conducted descriptively and relationally by mapping publication trends, author productivity, journal sources, countries, institutions, dominant keywords, co-authorship, co-occurrence, and conceptual clusters. The SLR analysis was conducted through thematic synthesis to classify the literature into themes such as automated scoring, automated feedback, multimodal assessment, learning analytics, metric validity, fairness, interpretability, and human oversight.

This strategy is consistent with the characteristics of the literature, which indicate that AI-based assessment has evolved from technical automation toward adaptive, multimodal, rubric-based measurement systems that require validation through human assessment [1], [3], [5], [7]. In addition, the analysis procedure was directed toward capturing key methodological issues, including the limitations of automated metrics, the need for human validation, the risk of bias, and the importance of multidimensional rubrics, as discussed in previous literature [10], [9], [13], [11], [12]. This approach can be replicated by other researchers using the same database, search date, search string, screening stages, inclusion and exclusion criteria, and analytical categories.

RESULTS AND DISCUSSION

Results

Development of Publications, Temporal Trends, Scientific Sources, Authors, Institutions, Countries, and Scientific Collaboration Patterns

The results of the literature selection indicate that studies on measurement innovation in Artificial Intelligence-based educational assessment have developed into an emerging and increasingly prominent research theme during the 2021–2026 period. Based on the article selection flow, the initial search conducted through the Scopus database on May 8, 2026, yielded 2,691 documents. The search strategy was applied to article titles, abstracts, and keywords using a combination of the terms “AI Assessment,” “Educational Measurement,” and “Automated Evaluation.” After several screening stages, 107 articles were included in the review. The publication trend shows a consistent increase in the number of publications from year to year, as presented in the graph below.

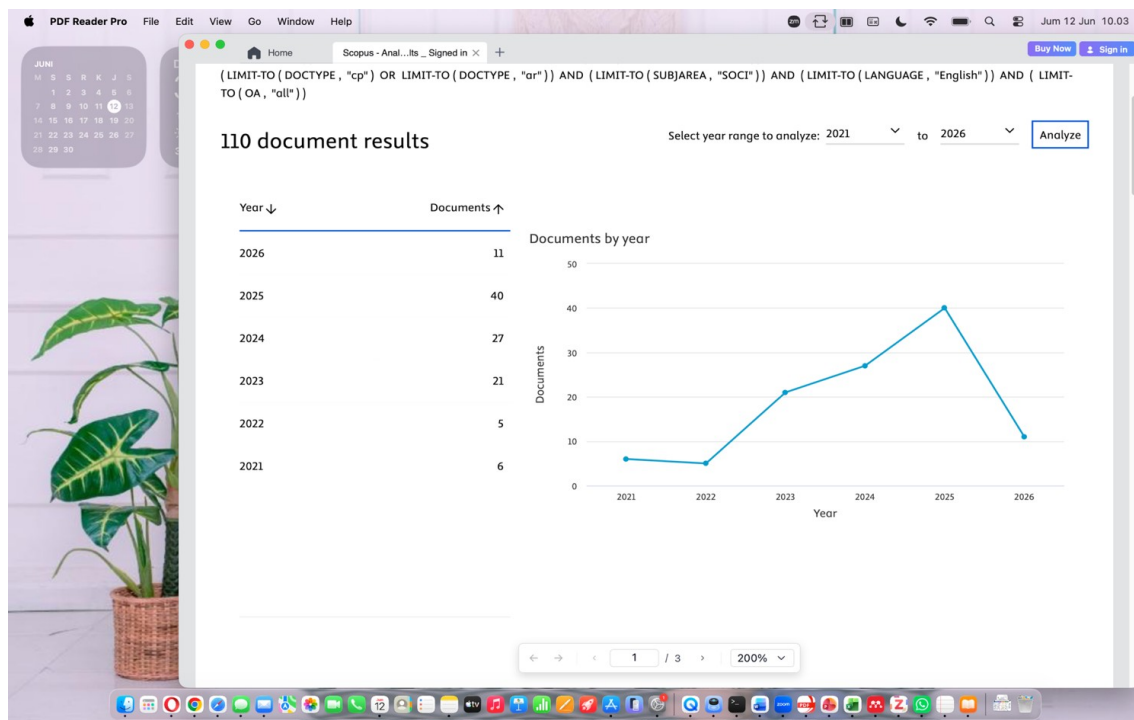


Figure 2. Grafik Documents by Year

The Documents by Year graph shows that research on measurement innovation in Artificial Intelligence-based educational assessment experienced strong growth during the 2023–2025 period. In 2021, there were 6 documents, followed by a slight decrease to 5 documents in 2022. A sharp increase began in 2023, with 21 documents, and continued in 2024, reaching 27 documents. The highest number of publications occurred in 2025, with 40 documents, while 2026 recorded 11 documents because the data were collected only up to May 8, 2026. This pattern indicates that AI assessment research has moved from a relatively limited initial issue toward a more established and widely discussed research theme.

The increase in publications after 2023 can be understood as a consequence of the development of large language models, generative AI, automated scoring, and learning analytics, which have begun to be widely used in education. In the early phase, AI was mostly positioned as a learning support tool or an administrative support system. However, after the development of generative models, scholarly attention began to shift toward the ability of AI to assess responses, provide feedback, identify learning patterns, and support academic decision-making. This shift contributed to the sharp increase in the number of documents from 2023 to 2025. The apparent decline in 2026 does not indicate a weakening of research interest, because the publication year had not yet been completed when the data were collected. The increase in the number of publications from 2023 to 2025 can also be observed in the Documents per Year by Source graph below.

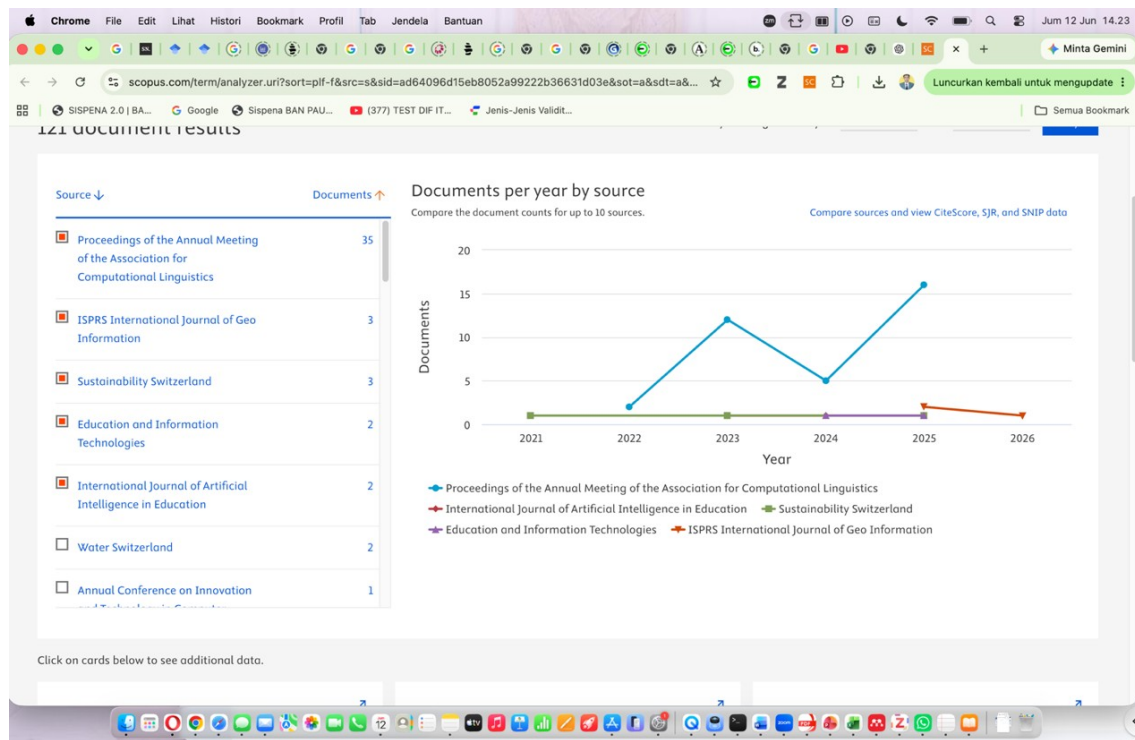


Figure 3. Grafik Documents per Year by Source

The Documents per Year by Source graph shows that the most dominant source is the Proceedings of the Annual Meeting of the Association for Computational Linguistics, with 35 documents. This dominance indicates that the development of AI-based assessment is strongly influenced by natural language processing, computational linguistics, and large language models. Other sources, such as ISPRS International Journal of Geo-Information, Sustainability Switzerland, Education and Information Technologies, and the International Journal of Artificial Intelligence in Education, contributed fewer documents, but they demonstrate the expansion of this research area into digital education, geospatial studies, sustainability, and learning technologies. This pattern occurs because AI assessment is no longer limited to written tests but has expanded to the evaluation of text, images, code, diagrams, spatial data, and digital performance. The dominance of computational linguistics forums also indicates that the technical aspects of AI remain more prominent than psychometric and pedagogical studies. The Documents per Year by Source graph presented above shows an increase in publications across these publication venues. This increase is also reflected in the number of authors who published articles in these venues, as shown in the Documents by Author graph.

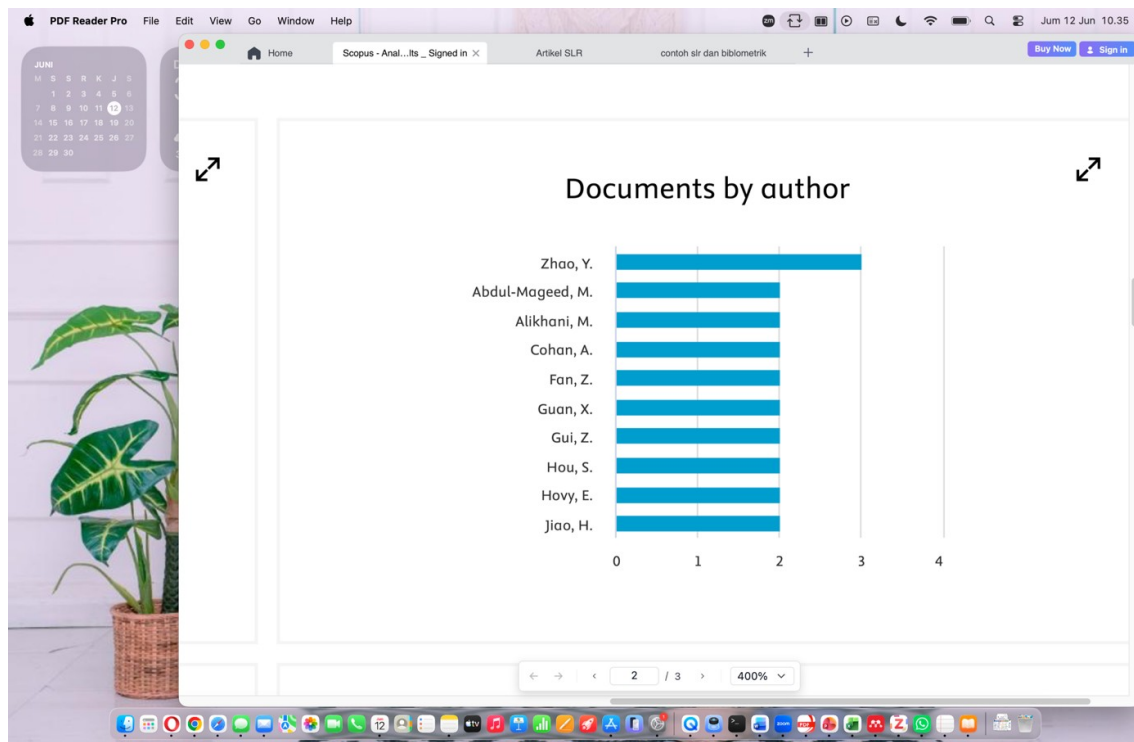


Figure 4. Grafik Documents by Author

The Documents by Author graph shows that author productivity is relatively dispersed and has not yet been concentrated around a single dominant scholar. Zhao, Y. is the most productive author, with 3 documents. Other authors, such as Abdul-Mageed, M.; Alikhani, M.; Cohan, A.; Fan, Z.; Guan, X.; Gui, Z.; Hou, S.; Hovy, E.; and Jiao, H., each contributed approximately 2 documents. This pattern indicates that research on AI-based measurement innovation is still in a growth phase, with many academic groups contributing in parallel. This field has not yet formed a single dominant research school; instead, it is developing through multidisciplinary networks.

The dispersion of authors occurs because AI-based assessment involves multiple areas of expertise. Research in this area requires not only experts in education or measurement but also specialists in machine learning, natural language processing, data science, computational linguistics, and learning technologies. In this context, productive authors are generally positioned at the intersection of AI model development and automated evaluation. This phenomenon explains why publications are often produced by collaborative teams rather than by individual researchers. AI assessment research requires the integration of technical model capabilities and assessment validity principles. Many publications in this field are produced by collaborative teams from countries with substantial research funding, as shown in the Documents by Country/Territory graph below.

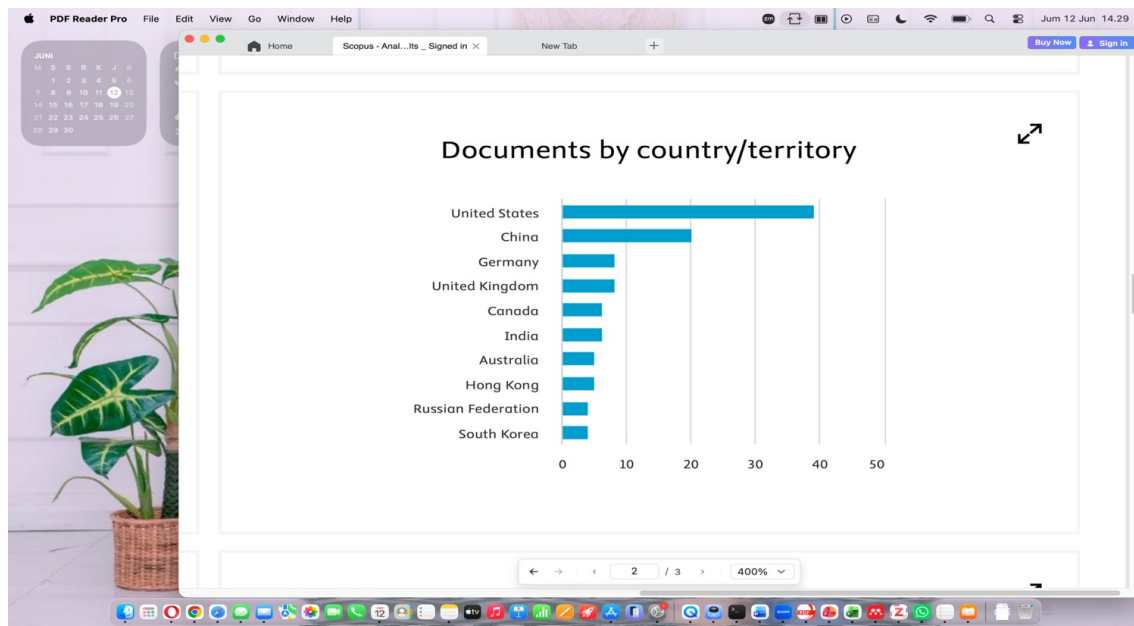


Figure 5. Grafik Documents by Country/Territory

The Documents by Country/Territory graph shows the dominance of the United States, with approximately 39 documents, followed by China, with approximately 20 documents. Other countries, such as Germany and the United Kingdom, contributed around 8 documents, while Canada and India contributed around 6 documents. Australia and Hong Kong contributed approximately 5 documents, whereas the Russian Federation and South Korea each contributed around 4 documents. This pattern indicates that knowledge production on AI-based assessment is still largely driven by countries with strong technological research ecosystems. The dominance of the United States and China is related to their strong funding capacity, computing infrastructure, AI laboratories, and extensive international publication networks. The dominant role of major countries is also reflected in the Documents by Funding Sponsor graph below.



Figure 6. Grafik Documents by Funding Sponsor

The Documents by Funding Sponsor graph shows that the National Science Foundation is the most dominant sponsor, with approximately 8 documents, followed by the National Natural Science Foundation of China, with approximately 4 documents. Other sponsors, such as federal ministries, national research foundations, the Russian Science Foundation, the Canada Foundation, Canada Research Chairs, the Defense Advanced Research Projects Agency, the Engineering and Physical Sciences Research Council, and the Intelligence Advanced Research Projects Activity, contributed approximately 2–3 documents. This pattern shows that AI assessment research is not only an academic issue but also part of national strategic agendas in the development of technology, data, and artificial intelligence. Such funding support exists

because AI is viewed as having economic, scientific, educational, and geopolitical value. This finding confirms that the direction of AI assessment research is strongly influenced by funding power, institutional capacity, and national technology priorities. The presence of major funding sponsors is also reflected in institutional affiliations. This can be observed in the Documents by Affiliation graph, which shows that Carnegie Mellon University in the United States is the most productive institution, with approximately 6 documents.

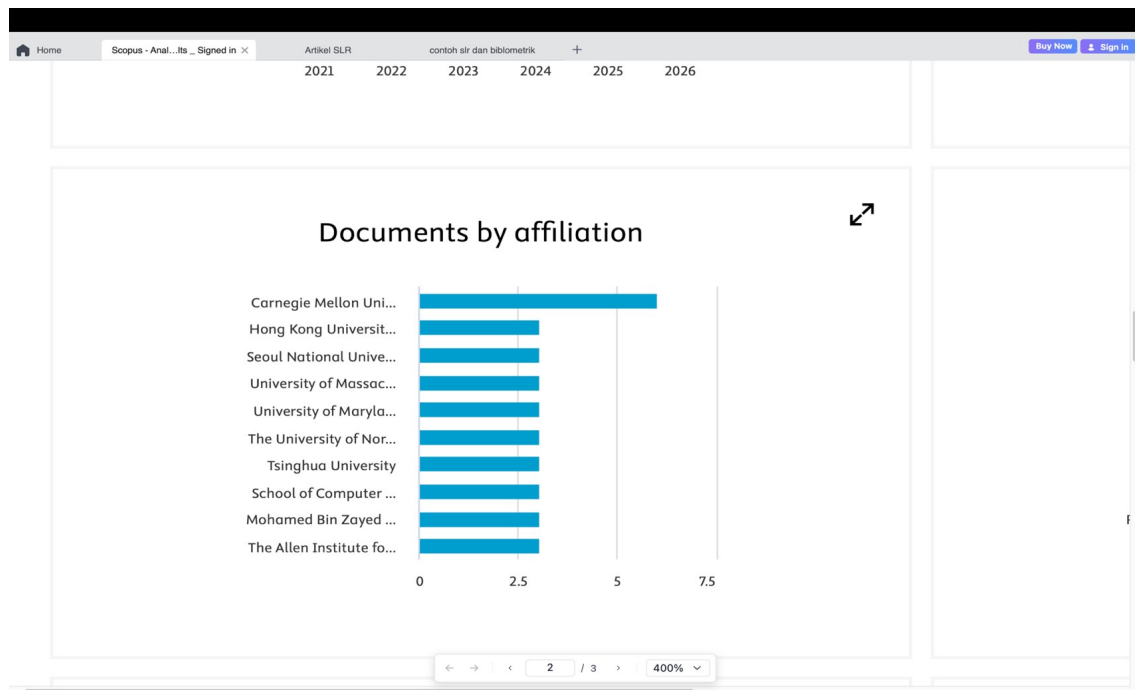


Figure 7. Grafik Documents by Funding Affiliation

In the graph above, other institutions, such as Hong Kong University, Seoul National University, the University of Massachusetts, the University of Maryland, The University of North Carolina, Tsinghua University, the School of Computer Science, Mohamed Bin Zayed University, and The Allen Institute for AI, each contributed approximately 3 documents. This pattern indicates that institutional contributions mainly come from universities and research centers with strong capacities in computer science, artificial intelligence, and educational technology. This occurs because the development of AI-based assessment requires computing resources, datasets, language models, and interdisciplinary collaboration. Institutions with advanced research infrastructure are more capable of producing evaluation models, benchmarks, and international publications.

However, the dominance of technology-oriented institutions also provides a critical note for the future direction of AI assessment development. Strong computing capacity and academic reputation do not automatically guarantee that the resulting assessment systems are pedagogically valid. AI systems may demonstrate strong technical performance, but they still require rubrics, expert validation, bias control, score interpretability, and alignment with learning objectives. This explains why AI assessment research needs to move beyond automated evaluation toward trustworthy AI-based educational measurement. The institutional affiliations shown in the graph indicate that this research area is still largely driven by technology centers; therefore, contributions from educational institutions, schools, and developing countries need to be expanded.

Themes, Conceptual Clusters, Dominant Keywords, and the Direction of Knowledge Evolution

Main Themes and Dominant Keywords

The VOSviewer visualization shows that the central themes in research on measurement innovation in Artificial Intelligence-based educational assessment are centered on the keywords artificial intelligence, learning systems, and language model. This can be seen in the Keyword Rank table below.

Table 1. Keyword Rank and Estimated Network Dominance

Rank	Keyword	Estimated Network Dominance*
1	Artificial Intelligence	Very High
2	Learning Systems	Very High
3	Language Model	Very High
4	Teaching	High
5	Natural Language Processing System	High

Rank	Keyword	Estimated Network Dominance*
6	Large Language Models	High
7	Automated Evaluation	High
8	Retrieval-Augmented Generation	Medium-High
9	Feedback	Medium-High
10	Education Computing	Medium
11	Benchmarking	Medium
12	Human Rater	Medium
13	AI-Augmented Learning	Medium
14	Modeling Performance	Medium
15	Computational Linguistics	Medium

These three keywords have the largest node sizes and occupy strategic positions in the network, indicating a high level of connection with other keywords. This condition suggests that the development of AI-based educational assessment is no longer focused solely on AI technology in general, but has shifted toward the application of AI as a system that supports evaluation, learning, and educational decision-making processes. The large node size of artificial intelligence indicates that AI functions as an umbrella concept connecting various modern assessment approaches. Meanwhile, learning systems and language model show that the current research focus is increasingly directed toward the use of intelligent learning systems and generative language models.

The dominance of keywords such as large language models, natural language processing system, and automated evaluation shows that AI-based measurement innovation is strongly influenced by advances in natural language processing technology. This phenomenon occurs because most educational assessment activities generate text-based data, such as essay responses, online discussions, learning reflections, and other academic assignments. Modern language models enable the evaluation process to be conducted automatically with increasingly high levels of accuracy. Therefore, many studies have begun to develop assessment systems capable of providing scores, feedback, and recommendations in real time. This development explains why keywords related to language model and automated evaluation have become dominant themes in the bibliometric network.

The presence of keywords such as feedback, human rater, and retrieval-augmented generation indicates that research attention is not focused solely on assessment automation. Recent literature has begun to question how the quality of AI-generated assessment results compares with human assessment. This occurs because the increasing capability of AI has not fully eliminated the need for human validation. Thus, the main research theme is currently moving toward the development of assessment systems that can combine technological efficiency with evaluation quality that is academically accountable.

Conceptual Clusters and Knowledge Structure

The VOSviewer map shows at least six main conceptual clusters that form the knowledge structure of this field. The first cluster is marked in red and centers on learning systems, education computing, algorithm analysis, and system features. This cluster represents research that focuses on the development of intelligent learning systems, algorithm design, and the technical characteristics of educational evaluation platforms. The dominance of this cluster indicates that the early foundation of AI-based assessment innovation still largely originates from the fields of computer science and educational technology. Researchers have sought to develop systems capable of automating the evaluation process with greater efficiency than conventional methods. This can be observed in the VOSviewer figure below.

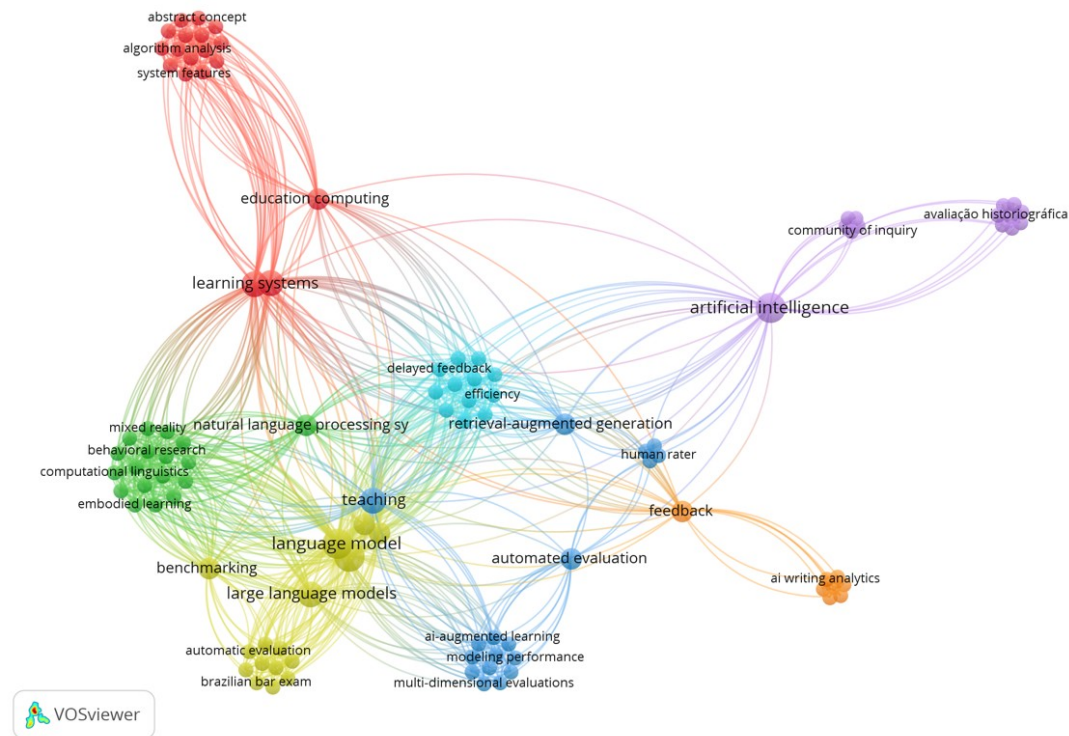


Figure 8. Network Keyword Visualization

The second cluster is marked in green and connects natural language processing system, computational linguistics, behavioral research, and embodied learning. This cluster shows that language-based evaluation has become one of the main centers of development in AI assessment. This occurs because most modern learning processes generate linguistic data that can be analyzed using NLP. The close relationship between NLP and computational linguistics indicates that the ability to understand human language is a key component in the development of adaptive assessment systems. In other words, the development of AI assessment depends heavily on advances in language processing technology.

The third and fourth clusters center on language model, large language models, automated evaluation, human rater, and retrieval-augmented generation. These clusters indicate the emergence of a new paradigm that uses generative language models as automated evaluators. The connection between automated evaluation and human rater suggests that the academic community has begun to pay greater attention to issues of validity, reliability, and the alignment of AI-generated assessment results with human judgment. Therefore, the direction of research is no longer limited to developing models that can assign scores, but also to ensuring that those scores are trustworthy and applicable in real educational contexts.

Direction of Knowledge Evolution

The network structure shows a fairly clear evolution of knowledge from a system-based approach toward a generative intelligence-based approach. In the early phase, research focused more on the development of learning systems, education computing, and algorithm analysis. The main focus during this period was how technology could support the evaluation process more quickly and efficiently. The systems developed were still oriented toward procedural automation and the processing of learning data. Therefore, research attention was directed more toward system performance than toward the quality of interpretation of assessment results.

The next development was marked by the increasing role of natural language processing system, computational linguistics, and language model. This shift occurred because digital learning environments generate an increasing amount of text-based data that requires more complex analytical techniques. AI no longer merely calculates scores, but begins to understand the meaning, context, and structure of students' responses. This evolution has enabled the emergence of more adaptive and personalized evaluation systems. As a result, research has begun to move from simple automation toward a deeper understanding of the learning process.

The most recent phase is reflected in the emergence of keywords such as large language models, retrieval-augmented generation, human rater, feedback, and AI-augmented learning. This cluster shows that the current research focus is moving toward the development of human-aligned, multimodal, and

validation-based assessment. AI is no longer viewed as a replacement for human evaluators, but rather as a partner in the educational measurement process. This development has occurred because increasing evidence shows that assessment quality is determined not only by technical accuracy, but also by the validity, fairness, interpretability, and transparency of evaluation results. Thus, the direction of knowledge evolution in this field is moving from automated scoring toward trustworthy AI-based educational measurement, namely an educational measurement system that integrates technological sophistication with valid and responsible assessment principles.

Forms, Functions, and Models of Artificial Intelligence-Based Measurement Innovation in Educational Assessment

Artificial Intelligence-based measurement innovation in educational assessment has developed in various complementary forms. The most dominant form is automated item scoring, which includes the evaluation of multiple-choice questions and item quality. The main functions of this innovation are to accelerate the assessment process, ensure consistency, and enable large-scale data processing. In addition, automated scoring can reduce dependence on human raters and minimize subjectivity. This phenomenon occurs because the large number of students and test items in modern education demands efficiency and speed in evaluation. Therefore, AI has become a solution for expanding assessment capacity while maintaining the reliability of scoring. This can be seen in the figure below.

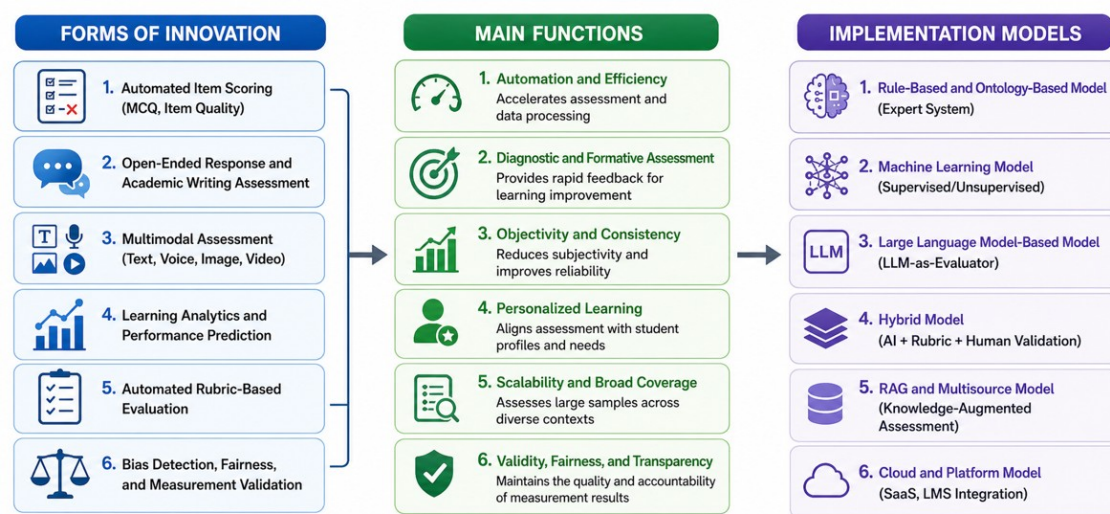


Figure 9. Forms, Functions, and Models of AI-Based Measurement Innovations in Assessment

In addition, open-ended response and academic writing assessment enable AI to automatically evaluate essays, short answers, and the quality of students' arguments. Its main function is to provide rapid formative feedback, help students understand their weaknesses, and improve their self-reflection skills. The models used typically rely on Large Language Models (LLMs) and Natural Language Processing (NLP) techniques, which are capable of analyzing grammar, coherence, and the logical structure of arguments. This phenomenon has emerged because evaluating free-text responses is more complex than evaluating multiple-choice items. Therefore, AI is needed to support scoring consistency while still incorporating human validation. In other words, this system combines machine efficiency with relevant pedagogical accuracy.

Another increasingly prominent innovation is multimodal assessment, which uses the integration of text, voice, images, and video to assess student performance more authentically. The main function of this form is to capture students' abilities more comprehensively, accommodate different learning styles, and enable the measurement of creativity and procedural skills. The models applied include vision-language models, AI-augmented learning, and retrieval-augmented generation (RAG), which integrates external data as a basis for evaluation. This phenomenon occurs because modern learning does not only produce written responses, but also complex digital artifacts. Therefore, text-based evaluation alone is not sufficient to assess competence comprehensively.

Furthermore, learning analytics and performance prediction are used to map learning patterns, risks of failure, and the effectiveness of student interactions. The main function of this form is to support personalized learning and data-driven decision-making. The models used include machine learning, predictive modeling, and adaptive analytics frameworks that allow educators to understand student profiles. This has emerged because digital education generates large amounts of interactive data, and the use of this data enables AI systems to provide adaptive evaluations that respond to individual abilities. This

combination supports the development of an assessment ecosystem that is adaptive, efficient, and grounded in validity.

Finally, automated rubric-based evaluation and bias detection ensure that AI-generated assessment results are aligned with pedagogical and ethical standards. The main function of this form is to maintain fairness, transparency, and score reliability. The implementation models usually combine AI with human validation, hybrid rubric-based assessment, or cloud-scalable platform models for LMS integration. This phenomenon occurs because, although AI can perform assessment quickly, aspects of validity, fairness, and accountability still require human oversight. Therefore, Artificial Intelligence-based measurement innovation in educational assessment is currently moving beyond technical automation toward evaluation systems that are human-aligned, multimodal, and adaptive.

Issues of Validity, Reliability, Fairness, Interpretability, Human Oversight, and Alignment with Human Assessment

Validity is a central issue in AI-based assessment because the scores generated by the system do not automatically reflect the pedagogical construct intended to be measured. AI can provide fast and consistent evaluation; however, its results depend heavily on rubrics, datasets, and educational contexts. This can be seen in the table of critical dimensions in AI assessment and their implications below.

Table 2. Critical Dimensions in AI Assessment and Their Implications

Dimension	Definition / Focus	Research Implications	Sample Studies
Validity	The extent to which AI-based measurement accurately reflects the construct being measured	AI needs to be compared with human judgment, rubrics, and pedagogical contexts	Jiang et al. [9], He and Ghassemiazghandi [10]
Reliability	The consistency of AI-generated scores when tested repeatedly	Models need to be tested using different datasets, consistent annotations, and appropriate metrics	Zhao et al. [33], Fan et al. [25]
Fairness	Bias-free assessment across gender, language, culture, or response position	Requires algorithmic adjustment and data diversification	H. Li et al. [34], Panfilova and Valueva [35]
Interpretability	The ability of AI to explain scores or recommendations	Scores must be traceable through rubrics, rationales, or human-in-the-loop mechanisms	J.-W. Chen et al. [36], K. Xu et al. [30]
Human Oversight	Human involvement in validating AI-generated scores	Ensures that AI decisions are aligned with pedagogical standards	Y. V. He et al. [37], Shahzad and Abbas [8]
Human Alignment	The alignment of AI-generated scores with human assessment	Avoids model overfitting to automated metrics and maintains pedagogical accuracy	Pires et al. [28], Naismith et al. [6]

This phenomenon occurs because AI models, including Large Language Models and machine learning algorithms, operate based on historical and statistical data patterns rather than a deep conceptual understanding. As a result, AI-generated scores without human validation may reflect only surface-level similarities in responses rather than the quality of students' understanding or arguments. Therefore, research emphasizes the importance of comparing AI-generated scores with human judgment and validating rubrics to ensure that assessment results remain relevant and valid.

Reliability is also a major concern because the consistency of AI-generated scores is not always stable when retested on different datasets or with variations in annotation. The same model may produce different scores if the input or evaluation conditions change. This occurs because AI is probabilistic and highly sensitive to data structure and training parameters. To address this issue, studies recommend model calibration, repeated testing, and metric adjustment so that evaluation results can be consistently reliable. High reliability is important to ensure consistent and fair academic decisions and to minimize uncertainty in score-based decision-making using AI.

Fairness and interpretability are closely related dimensions. Fairness requires that AI-based assessment be free from bias related to factors such as gender, language, culture, or response position. Bias may emerge when models are trained on unrepresentative datasets, causing students to receive different scores even when their actual abilities are equivalent. Interpretability is crucial so that scores can be traced and explained, for example through rubrics, rationales, or human-in-the-loop mechanisms. This phenomenon occurs because many modern models operate as black boxes; therefore, without transparency, scores are difficult to justify. Explainable AI approaches and rubric validation help ensure that evaluation results are pedagogically and ethically trustworthy.

Human oversight and human alignment emphasize the role of humans in verifying and adjusting AI-generated assessment results so that they remain consistent with educational standards. Human oversight ensures that AI decisions align with pedagogical objectives, while human alignment prevents overfitting to automated metrics and maintains evaluation quality. This phenomenon occurs because strong technical performance in AI does not automatically guarantee valid and fair decisions. AI-generated scores need to be verified so that they reflect students' actual understanding. The integration of validity, reliability, fairness, interpretability, and human oversight forms a framework for AI assessment that is reliable, accountable, and human-aligned. Thus, AI-based measurement innovation becomes effective and relevant in the context of modern education.

Research Gap, Novelty Potential, and Future Research Agenda

The literature analysis shows that although the use of AI in educational assessment has developed rapidly, significant research gaps still need to be addressed. One of the main gaps is the limited construct validity of AI across different contexts, cultures, and student populations. This can be seen in the conceptual flow diagram of the research gap, novelty potential, future research agenda, and expected impact below.

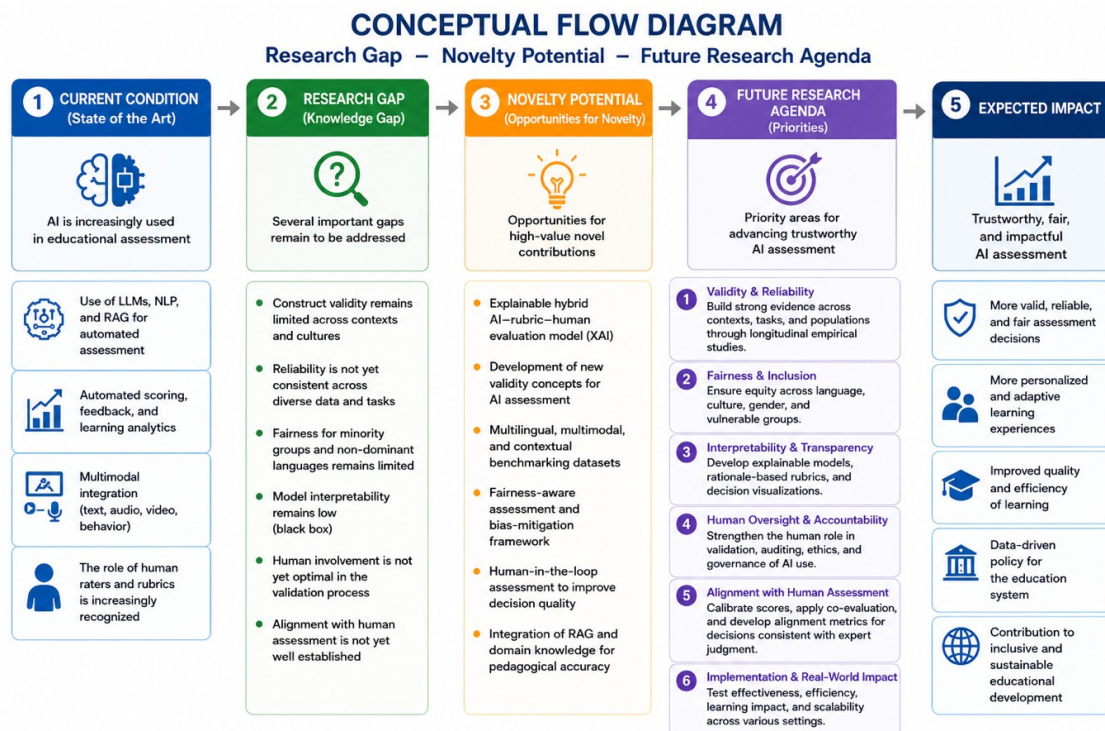


Figure 10. Conceptual Flow Diagram

Current models are often trained on homogeneous datasets, so their evaluation results do not fully represent students' abilities universally. Score reliability is also not yet consistent when applied to diverse data and tasks. In addition, fairness for minority groups and non-dominant languages remains limited, while model interpretability is low because of the black-box nature of algorithms. Human involvement in validating AI-generated scores is also not yet optimal, so alignment with human assessment has not been firmly established. This phenomenon occurs because early research focused more on technical accuracy and automation efficiency, while pedagogical, ethical, and contextual aspects have only recently begun to receive greater attention.

In terms of novelty potential, there are major opportunities to develop hybrid evaluation models that combine AI and human rubrics. This innovation enables assessment that is not only fast and consistent but also explainable and traceable through explainable AI (XAI). The development of new validity concepts and multilingual and multimodal benchmark datasets is also important for adapting evaluation systems to diverse learning contexts. In addition, the application of fairness-aware assessment frameworks and bias-mitigation strategies can help ensure that the scores generated are fair for all students. This phenomenon occurs because, although AI technology continues to develop, the integration of technical performance with valid pedagogical principles remains limited. Therefore, there is still substantial room for high-value innovation.

Another promising innovation is the use of human-in-the-loop assessment, which allows human evaluators to review, validate, and refine AI-generated results. This approach not only improves the quality of decisions but also maintains trust in the assessment system. The integration of retrieval-augmented generation (RAG) with domain knowledge enables AI systems to provide more accurate and pedagogically relevant responses. This phenomenon occurs because fully automated AI still has limitations in understanding the context and nuances of students' responses. Therefore, human intervention is needed to ensure that the resulting scores are valid and accountable.

Overall, the research gaps and novelty potential indicate a clear direction for future research: the focus should be placed on developing AI assessment that is valid, reliable, fair, interpretable, and human-aligned. Existing gaps provide a strong rationale for exploring hybrid models, fairness-aware frameworks, and multimodal datasets. Novelty potential emerges because the combination of AI technology with human oversight, RAG integration, and the development of new rubrics enables more adaptive, contextual, and pedagogically grounded innovation. This phenomenon reflects that although AI has brought major advances to assessment, further research is still needed to bridge the gap between technical efficiency and valid educational evaluation quality.

Discussion

Development of Publications, Temporal Trends, Sources, Authors, Institutions, Countries, and Scientific Collaboration Patterns

Research on measurement innovation in Artificial Intelligence (AI)-based educational assessment has experienced strong growth, particularly during the 2023–2026 period. The reviewed literature shows that this topic has developed at the intersection of educational assessment, artificial intelligence, machine learning, natural language processing, large language models, multimodal learning, automated scoring, and human–AI evaluation. The dominance of publications in the most recent period indicates that this theme is no longer a peripheral discourse, but has become a new center of attention in educational technology and academic measurement.

Ramesh et al. [7], [9], [11], [2], [6], [1], and [18] show that academic discussion has shifted from AI as a support tool toward AI as an instrument for evaluation and measurement. This trend is further strengthened by the studies of [3], [38], [33], [28], [37], [16], [14], [8], and [21]. Thus, the bibliometric map indicates that this field has accelerated following the development of Large Language Models (LLMs), GenAI, and multimodal Artificial Intelligence (AI).

The significance of this finding lies in the evidence that AI-based assessment studies are interdisciplinary. The reviewed literature shows contributions from education, computational linguistics, computer science, medicine, law, geospatial studies, digital humanities, and educational technology. [7] focus on the context of introductory programming education, [9] on translation evaluation, [10] on machine translation evaluation, Ostheimer et al. [11] on the validation of Natural Language Processing (NLP) metrics, and [29] on the accessibility evaluation of Massive Open Online Courses (MOOCs).

This diversity explains why AI measurement innovation cannot be limited to classrooms, examinations, or conventional rubrics. AI has entered various forms of knowledge representation, ranging from text, code, diagrams, reflection, and communication to multimodal artifacts. This finding is important for the field of educational assessment because it shows that measurement constructs need to be expanded in order to capture increasingly diverse forms of learning evidence. In terms of scientific collaboration, the findings show that this research area has developed through multidisciplinary clusters. In the domain of automated writing evaluation, [4], [8], [6], [39], and [4] form a cluster focused on academic writing assessment, coherence, reflection, and feedback quality. In the domain of LLM-as-judge and validation, [33], [28], [34], [10], [25], [40], and [32], [41] focus on reliability, alignment, bias, and rubric quality. In the multimodal domain, [15], [16], [14], [42], [43], and [44] show a collaborative direction among computer vision, multimodal Large Language Models (LLMs), performance evaluation, and artifact-based assessment. This pattern is important because it shows that AI-based educational measurement does not emerge from a single discipline, but from the convergence of assessment theory, data, algorithms, language, and learning domains.

Main Themes, Conceptual Clusters, Dominant Keywords, and the Direction of Knowledge Evolution

The reviewed literature forms at least five main conceptual clusters. The first cluster is automatic question generation and item quality evaluation, as shown in [7], [5], [3], and [1]. This cluster discusses AI's ability to evaluate the quality of multiple-choice questions, generate question stems, develop distractors, and classify test items based on Bloom's cognitive levels. The dominant keywords in this cluster include multiple-choice question, distractor generation, Bloom's Taxonomy, rubric-based

assessment, automated evaluation, item quality, and learning objectives. This finding shows that AI innovation in assessment initially developed from the need to improve the quality of assessment instruments. This occurs because multiple-choice questions and objective instruments remain widely used forms of assessment, but their quality is often influenced by the accuracy of indicators, distractors, and cognitive levels.

The second cluster is automated writing evaluation and free-text assessment. Bai & Stede [2], [39], [4], [8], and [6] show that AI can assess essays, academic writing, open-ended responses, discourse coherence, reflection, and feedback quality. This cluster is important because open-ended responses are more closely related to higher-order thinking skills than objective questions. However, assessing open-ended responses is also more difficult because it involves argument structure, content quality, context, language, and meaning. Therefore, the emergence of Natural Language Processing (NLP), Large Language Models (LLMs), and hybrid models in this cluster explains the need in modern education for assessment that is faster while remaining meaningful. Its significance lies in the fact that AI is beginning to open opportunities to expand formative assessment, which was previously limited by human workload.

The third cluster is automated metric validation and human judgment alignment. [9], [10], [11], [44], [13], [25], and [45] emphasize that automated scores should not be accepted without validation against human assessment. This cluster shows an epistemic shift from “AI can assess” toward “whether AI assessment can be trusted.” This shift is important because educational assessment is not only concerned with efficiency, but also with the consequences of academic decisions. If automated scores are not aligned with human judgment, measurement results may lead to diagnostic errors, biased decisions, or unfairness for students. Thus, the findings in this cluster strengthen the position of this study that human validation remains a core component of AI-based assessment.

The fourth cluster is multimodal, authentic, and performance-based assessment, while the fifth cluster is learning analytics, formative assessment, gamification, and digital learning ecosystems. [15], [16], [14], [46], [47], [42], [43], and [48] show that AI-based measurement objects have expanded from text to images, diagrams, videos, code, creativity, multimodal responses, and procedural performance. [18], [21], [23], [38], [20], [17] show that AI is also used to monitor learning processes, real-time evaluation, discussion forums, project-based learning, gamified learning, peer assessment, and the calibration of prospective teachers’ assessment.

The learning analytics cluster shows that assessment is no longer limited to final outcomes, but has entered the monitoring of learning processes, interactions, participation, motivation, and learning performance, as shown by [18], [21], [23], [38], [20], [17], and [19]. This finding is important because it confirms that AI opens the possibility of more granular, continuous, and data-driven measurement. This direction of knowledge evolution shows a shift from assessment as the correction of final results toward assessment as a system for monitoring learning processes. This is important because digital learning produces data traces that cannot be adequately interpreted through traditional instruments alone.

Forms, Functions, and Models of AI-Based Measurement Innovation in Educational Assessment

The earliest and most dominant form of innovation is automated scoring. Ramesh et al. [7] show that Large Language Models (LLMs) can be used to assess the quality of multiple-choice questions based on information sufficiency, clarity of the correct answer, uniqueness of options, distractor quality, and code correctness. Ahmed & Sorour [1] show that intelligent systems can evaluate the quality of examination papers through formal specifications and the classification of Bloom’s cognitive levels. Bai & Stede [2] map automated free-text evaluation through feature-based, neural, and hybrid approaches. H. Chen & Pan [4] show that automated evaluation scoring can accelerate writing assessment, but has not replaced instructors in assessing structure, logic, and argumentation. Naismith et al. [6] show that GPT-4 can assess discourse coherence and provide rationales to improve score interpretability. J.-W. Chen et al. [36] show that Natural Language Processing (NLP) can classify the quality of reflection and feedback in workplace-based assessment.

The main function of automated scoring is not only to accelerate grading, but also to expand measurement capacity. This occurs because the number of students, responses, and learning artifacts in digital education continues to increase. Under these conditions, manual assessment finds it difficult to maintain speed, consistency, and scalability. However, the reviewed literature consistently shows that automation is not sufficient without rubrics and validation. Ramesh et al. [7] emphasize the importance of gold-label quality and annotation consistency, while H. Chen & Pan [4] show that automated assessment remains weak in interpreting complex argument structures. Therefore, automated scoring in this study is understood not as a replacement for human raters, but as a support system that needs to be controlled through constructs, rubrics, and validation.

The second form of innovation is automated feedback and formative AI assessment. Shahzad & Abbas [8] show that AI writing analytics can provide rapid feedback on grammar, vocabulary, coherence, and revision support, but remains limited in assessing argumentation, creativity, academic style, cultural sensitivity, and content quality. J.-W. Chen et al. [36] show that NLP can help monitor the quality of reflection and feedback in competency-based assessment, but is not intended to replace expert review. Yoshida et al. [21] show that AI in online discussion forums can function as a virtual teacher, peer/message matcher, automated evaluator, and feedback provider. This finding is important because formative assessment requires rapid feedback so that students can improve their learning processes. Thus, the function of AI develops from score generation into support for learning regulation and instructional improvement.

The third form of innovation is multimodal assessment and learning analytics. Rahmanian et al. [16] show that multimodal LLMs can be used to evaluate students' ER diagrams, although they still face difficulties with cardinality and complex visual symbols. Panfilova and Valueva [35] show the use of deep learning and explainable AI for creativity scoring, while still facing issues of interpretability and cultural validity. Bang Andersen et al. [14] extend assessment to video-based micro-assessment of procedural skills. Wu et al. [46], show that geospatial code evaluation can be designed as domain-specific, multimodal, and unit-test-based. These findings show that the objects of educational measurement are becoming increasingly broad, so assessment models need to move from textual responses toward authentic, visual, temporal, and performative evidence of learning.

Issues of Validity, Reliability, Fairness, Interpretability, Human Oversight, and Alignment with Human Assessment

Measurement quality issues are at the core of discussions on Artificial Intelligence (AI) assessment. L. He & Ghassemiazghandi [10] show that BLEU has limitations because it only captures surface-level lexical similarity and does not sufficiently capture nuances of meaning, culture, ideology, fluency, and coherence. Ostheimer et al. [11] show that many studies on text style transfer use automated metrics without adequate validation. Wang and Li [12], [44] found that automated metrics may not align with human evaluation. Jon & Bojar [13] show the risk of metric gaming, a condition in which systems can optimize metric scores without guaranteeing the substantive quality of the output. This finding is important because, in educational assessment, metrics must not become an end in themselves. Metrics must serve as valid representations of the constructs being measured.

Reliability is also an important issue. Ramesh et al. [7], Zhao et al. [33], Naismith et al. [6], and H. Chen & Pan [4] reveal issues related to annotation consistency, gaps with experts, limited reproducibility, and classification accuracy that differs across levels. This means that AI models may appear strong in one context, but their performance may not necessarily remain stable in other contexts. Fairness and bias are evident in H. Li et al. [34], who show selection bias, position bias, and token bias in LLM-as-judge. Panfilova and Valueva [35] also show the potential for visual bias in creativity scoring. This finding is important because educational assessment must ensure equal treatment for students from different linguistic, cultural, and experiential backgrounds. If AI models are trained on unrepresentative data, the resulting scores may reproduce inequality.

Interpretability is the next key issue. Naismith et al. [6], Fan et al. [25], [40], and L. He & Ghassemiazghandi [10] emphasize the importance of rationales, adaptive rubrics, multifaceted evaluation, faithfulness, contextual relevance, and disciplinary criteria. Interpretability is important because educational users need to understand why students receive certain scores. Without traceable explanations, AI-generated scores are difficult to justify academically and ethically. Human oversight is a central issue because Shahzad & Abbas [8], Choi et al. [22], Pires et al. [28], and L. He & Ghassemiazghandi [10] position AI as a support tool that still requires human validation, particularly in high-stakes assessment. Thus, alignment with human assessment becomes an indicator that AI innovation must not only be technically accurate, but also explainable, testable, calibrated, and pedagogically controlled.

The significance is substantial because it determines whether AI is appropriate for use in educational decision-making. Choi et al. [22] show that the detection of inauthentic text in high-stakes assessment still requires a human-in-the-loop approach. This shows that although AI can support academic integrity, final decisions still require human judgment. The practical implication is that AI assessment systems need to be designed with rubric audits, human validation, score rationale tracking, bias checks, and mechanisms for appeal or verification. The theoretical implication is that AI assessment validity must be viewed as system validity, not merely score validity. This means that what needs to be tested is not only the model output, but also the data, rubrics, context, interpretation, score use, and consequences of decisions. Adjacent automated-evaluation research has also examined uncertainty in language-model answers, Generative-AI-

assisted quality evaluation outside education, and transparent multi-LLM grading architectures, providing complementary comparisons for trustworthiness and reproducibility [49], [50], [51].

Research Gap, Novelty Potential, and Future Research Agenda

The main research gap lies in the absence of an integrated framework that comprehensively connects AI technological innovation with the principles of educational measurement. The reviewed literature remains fragmented by domain, such as Multiple Choice Question (MCQ) assessment, translation evaluation, automated writing evaluation, code evaluation, legal writing evaluation, multimodal assessment, clinical Question Answering (QA) evaluation, dialogue evaluation, and creativity scoring. Ramesh et al. [7], Kumar et al. [5], Byun & Choi [3], and Ahmed & Sorour [1] focus on instruments and test items. Bai & Stede [2], H. Chen & Pan [4], Shahzad & Abbas [8], Naismith et al. [6], and H. Chen & Pan [4] focus on writing, open-ended responses, and feedback. Kulkarni et al. [15], Rahmanian et al. [16], Bang Andersen et al. [14], and Nguyen & Park [43] focus on multimodality and performance. Zhao et al. [33], Pires et al. [28], H. Li et al. [34], L. He & Ghassemiazghandi [10], Fan et al. [25] and Zhang et al. [32], [41] focus on LLM-as-judge validation.

The most prominent methodological gap is the weak standardization of evaluation and the inconsistent alignment between automated metrics and human assessment. Ostheimer et al. [11] show that many studies on text style transfer use automated metrics without adequate evidence of validity and with limited reporting of statistical significance. Wang and Li [12], [44] found that automated metrics may not align with, and may even contradict, human evaluation. L. He & Ghassemiazghandi [10] show the limitations of Bilingual Evaluation Understudy (BLEU) as a single metric because it captures only surface-level lexical similarity. Jon & Bojar [13] show the risk of metric gaming when systems are optimized to satisfy particular metrics. H. Li et al. [34] show that LLM-as-judge can experience selection bias, position bias, and token bias. Panfilova and Valueva [35] show that creativity models may be biased toward certain visual elements. [29] further indicate that digital evaluation is also influenced by accessibility and sample representativeness.

The novelty potential of this study lies in developing a synthesis that places “measurement” at the center of analysis, rather than merely focusing on the use of AI in education. This novelty is important because it shifts the focus from technological questions to measurement questions, namely whether AI outputs are truly valid, reliable, fair, explainable, and aligned with human assessment. This study integrates bibliometric mapping, conceptual clusters, forms of measurement innovation, validation issues, and future research agendas into a single framework of trustworthy AI-based educational measurement. Its originality lies in combining automated scoring, automated feedback, multimodal assessment, learning analytics, validity, reliability, fairness, interpretability, and human oversight within one conceptual model. Thus, the contribution of this study is not only to classify the literature, but also to build a new conceptual direction for AI-based educational assessment.

The future research agenda should be directed toward six priorities. First, research needs to strengthen evidence of validity and reliability across contexts, tasks, educational levels, languages, and populations. Second, research needs to develop multilingual, multimodal, and contextual benchmark datasets so that models are not valid only on dominant data. Third, research needs to strengthen fairness-aware assessment and bias-mitigation frameworks. Fourth, research needs to develop explainable AI that can explain scores through rubrics, rationales, and decision visualizations. Fifth, research needs to strengthen human-in-the-loop assessment so that important decisions remain under human control. Sixth, research needs to examine the real impact of AI on learning, academic decisions, educators’ workload, and students’ experiences. This agenda is important because Artificial Intelligence (AI)-based measurement innovation is meaningful only if it can improve the quality of educational decisions, not merely accelerate technical processes.

Significance of the Research Findings

The findings of this study are significant because they shift the way Artificial Intelligence (AI) is understood in education. Many previous studies have demonstrated that AI can assess questions, texts, translations, code, diagrams, creativity, reflection, communication, chatbots, discussion forums, multimodal responses, and educational quality. However, the same studies also show consistent limitations, including model bias, domain specificity, limited generalization, misalignment with human judgment, weaknesses in automated metrics, and risks of unsupervised use. Therefore, this study is important because it brings together two dimensions that are often discussed separately: technological potential and the requirements for measurement legitimacy. This study emphasizes that AI can expand, accelerate, and enrich educational assessment, but its academic legitimacy still depends on validity, reliability, fairness, interpretability, and human control. In other words, the findings of this study show that measurement innovation does not mean

replacing humans with machines, but building assessment systems that are faster, richer in data, more adaptive, more transparent, and still human-centered.

Conceptually, this study contributes by developing a framework that understands AI assessment as a measurement system, not merely as a technological application. This contribution is important because the field of education is often tempted to view AI in terms of efficiency, automation, or model sophistication. This study shows that the more fundamental question is whether AI systems truly measure the appropriate construct, produce stable scores, operate fairly, can be explained, and align with human assessment. Through this framework, this study enriches the discussion of educational assessment through the concept of trustworthy AI-based educational measurement. This concept positions AI as a system that needs to be validated, audited, and calibrated, rather than as an independent authority in assessment. This conceptual contribution is relevant to the development of assessment theory in the digital era.

Methodologically, this study contributes through the use of a hybrid approach combining a systematic literature review and bibliometric analysis. This approach enables the study not only to map the number of publications, sources, authors, countries, and institutions, but also to synthesize the content of the literature, conceptual clusters, forms of innovation, validation issues, and future agendas. This approach is important because the AI assessment literature is broad and spread across domains. Without systematic synthesis, the findings of previous studies would remain fragmented and difficult to use in building a direction for educational assessment development. By combining quantitative mapping and qualitative synthesis, this study is able to show the structure of knowledge while also interpreting its scholarly meaning. Thus, the methodological contribution of this study lies in its ability to bridge bibliometric mapping with conceptual argumentation.

Practically, the findings of this study provide guidance for educators, system developers, educational institutions, and policymakers. Educators can use these findings to understand that AI should be used as a tool for assessment and feedback, not as a replacement for professional decision-making. System developers can use these findings to design assessment systems that are rubric-based, explainable, and equipped with human validation mechanisms. Educational institutions can use these findings to establish governance for the use of AI in assessment, including bias audits, score rationale tracking, and appeal procedures. Policymakers can use these findings as a basis for formulating minimum standards for the use of AI in educational decisions. These practical implications are important because the use of AI in high-stakes assessment can directly affect students' academic opportunities.

Research Implications

The theoretical implication of this study is the need to expand the concept of validity in Artificial Intelligence (AI)-based educational assessment. Validity should not be understood only as the relationship between scores and constructs, but should also include training data, rubric design, model mechanisms, output interpretability, alignment with human judgment, and the consequences of score use. The methodological implication is the need for standardization in AI assessment evaluation. Every AI assessment study should report the dataset, rubric, annotation procedures, reliability testing, human validation, bias checks, and the limits of model use. The pedagogical implication is that AI should be used to strengthen the learning process through formative feedback, learning analytics, and adaptive assessment. The ethical implication is that every use of AI in assessment must be accompanied by the principles of fairness, transparency, human oversight, and accountability. Thus, the implications of this study cover theory, method, practice, and governance.

Research Limitations

This study has several limitations. First, this review depends on the literature examined and the databases used in the selection process, so the findings may be influenced by search scope, search strings, inclusion criteria, and article availability. Second, most of the literature discussed comes from specific contexts, such as programming, translation, writing assessment, clinical Question Answering (QA), geospatial studies, diagrams, creativity, or discussion forums, so generalization to all forms of educational assessment should be made cautiously. Third, many of the reviewed studies are still focused on model development, benchmarking, or limited evaluation, so evidence of long-term impact on learning outcomes and educational decisions remains limited. Fourth, several findings related to fairness, interpretability, and human oversight are still at the conceptual stage or limited to domain-specific studies. Fifth, bibliometric analysis shows publication patterns and keyword relationships, but does not directly prove the pedagogical effectiveness of each AI model. Therefore, future research needs to conduct empirical validation across contexts, languages, educational levels, and student populations.

Overall, this study contributes to the field of educational assessment by emphasizing that the future of AI assessment must be built on valid, reliable, fair, transparent, and human-supervised measurement

principles. This finding is important because education requires not only fast technology, but also trustworthy evaluation systems. Artificial Intelligence (AI) can indeed expand assessment capacity, but its academic value emerges only when AI outputs can be evidenced, explained, and accounted for. Thus, this study provides a conceptual and methodological foundation for the development of AI-based educational assessment that is not only automated, but also meaningful, ethical, and human-aligned.

CONCLUSION

This study concludes that measurement innovation in Artificial Intelligence-based educational assessment has developed significantly from simple automated scoring models toward measurement systems that are more adaptive, multimodal, data-driven, and oriented toward human validation. The main findings show that AI has been used in various forms of assessment, such as automated scoring, automated feedback, multimodal assessment, learning analytics, item quality evaluation, open-ended response assessment, academic writing evaluation, code assessment, diagram evaluation, and performative assessment based on digital artifacts.

From a bibliometric perspective, this field shows strong development during the 2023–2026 period, with research increasingly focused on LLMs, GenAI, human-aligned validation, fairness, interpretability, and multimodality. From a thematic perspective, the literature forms several main clusters, including automatic question generation, automated writing evaluation, human judgment alignment, multimodal assessment, and learning analytics. These findings show that AI not only accelerates the assessment process, but also expands the objects, methods, and evidence of measurement in educational assessment.

The main contribution of this study lies in developing a conceptual synthesis that places “measurement” at the core of the analysis, rather than merely examining the use of AI as a learning support tool. Thus, this study contributes to the field of educational assessment by mapping the development of the literature, identifying conceptual clusters, emphasizing issues of validity, reliability, fairness, interpretability, and human oversight, and formulating a direction toward trustworthy AI-based educational measurement.

Recommendations for future research include the need to develop AI-based assessment models that are not only technically accurate, but also pedagogically valid, procedurally fair, and academically auditable. Future studies need to examine the validity and reliability of AI systems across different educational levels, fields of study, languages, response formats, and assessment contexts. Further studies also need to develop multidimensional rubrics capable of measuring response quality, argumentation, creativity, coherence, performance, and multimodal evidence of learning more comprehensively.

In addition, research is needed to specifically evaluate fairness, bias, selection bias, position bias, token bias, and the potential for metric gaming in AI-based scoring systems. The development of AI assessment also needs to apply human-in-the-loop principles, especially for high-impact educational decisions, so that assessment results can still be verified by educators, evaluators, or assessment experts. From a methodological perspective, future research needs to use more comprehensive bibliometric data to visualize co-authorship, co-citation, bibliographic coupling, co-word analysis, and thematic evolution more precisely. Thus, the next research agenda should be directed toward strengthening the integration of AI technology, educational measurement theory, human validation, ethical governance, and transparent and accountable assessment design.

REFERENCES

- [1] H. M. M. Ahmed and S. E. Sorour, “Classification-driven intelligent system for automated evaluation of higher education exam paper quality,” *Education and Information Technologies*, vol. 29, no. 15, pp. 19835–19861, 2024, doi: 10.1007/s10639-024-12555-9.
- [2] X. Bai and M. Stede, “A Survey of Current Machine Learning Approaches to Student Free-Text Evaluation for Intelligent Tutoring,” *International Journal of Artificial Intelligence in Education*, vol. 33, no. 4, pp. 992–1030, 2023, doi: 10.1007/s40593-022-00323-0.
- [3] G. Byun and J. D. Choi, “D-GEN: Automatic Distractor Generation and Evaluation for Reliable Assessment of Generative Models,” In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (pp. 3316–3349). Association for Computational Linguistics (ACL), 2025, doi: 10.18653/v1/2025.findings-acl.174.
- [4] H. Chen and J. Pan, “Computer or human: a comparative study of automated evaluation scoring and instructors’ feedback on Chinese college students’ English writing,” *Asian-Pacific Journal of Second and Foreign Language Education*, 7(1), 2022, doi: 10.1186/s40862-022-00171-4.
- [5] A. P. Kumar, A. Nayak, M. S. K., and K. Ghosh, “A Novel Framework for the Generation of Multiple Choice Question Stems Using Semantic and Machine-Learning Techniques,” *International Journal of Artificial Intelligence in Education*, vol. 34, no. 2, pp. 332–375, 2024, doi: 10.1007/s40593-023-00333-6.
- [6] B. Naismith, P. Mulcaire, and J. Burstein, “Automated Evaluation of Written Discourse Coherence Using GPT-4,” In E. Kochmar, J. Burstein, A. Horbach, A. Horbach, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z.

- Yuan, T. Zesch, & T. Zesch (Eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (pp. 394–403). Association for Computational Linguistics (ACL), 2023, doi: 10.18653/v1/2023.bea-1.32.
- [7] A. Ramesh, A. Agarwal, J. A. Doughty, K. Ramaneti, J. Savelka, and M. Sakr, “A Benchmark for Testing the Capabilities of LLMs in Assessing the Quality of Multiple-choice Questions in Introductory Programming Education,” *SIGCSE Virtual 2024 - Proceedings of the 2024 ACM Virtual Global Computing Education Conference V. 1*, 193–199, 2024, doi: 10.1145/3649165.3690123.
- [8] W. Shahzad and F. Abbas, “Automated Evaluation of ESL Learners’ English Writing Skills in English-Medium Instruction (EMI) through AI Writing Analytics,” *Journal of Communication, Language and Culture*, vol. 6, no. 1, pp. 147–166, 2026, doi: 10.33093/jclc.2026.6.1.8.
- [9] L. Jiang, Y. Jiang, and L. Han, “The potential of ChatGPT in translation evaluation: A case study of the Chinese-Portuguese machine translation,” *Cadernos de Tradução*, 44(1), 2024, doi: 10.5007/2175-7968.2024.e98613.
- [10] L. He and M. Ghassemiazghandi, “A BLEU-Based Evaluation of ChatGPT’s Chinese-to-English Translation,” *Theory and Practice in Language Studies*, vol. 15, no. 12, pp. 4017–4026, 2025, doi: 10.17507/tpls.1512.21.
- [11] P. Ostheimer, M. Nagda, M. Kloft, and S. Fellenz, “A Call for Standardization and Validation of Text Style Transfer Evaluation,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 10791–10815, 2023, doi: 10.18653/v1/2023.findings-acl.687.
- [12] Q. Wang and Y. Li, “Multi-Model Fusion for Street Visual Quality Evaluation,” *ISPRS International Journal of Geo-Information*, 15(4), 2026, doi: 10.3390/ijgi15040158.
- [13] J. Jon and O. Bojar, “Breeding Machine Translations: Evolutionary approach to survive and thrive in the world of automated evaluation,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1, 2191–2212, 2023, doi: 10.18653/v1/2023.acl-long.122.
- [14] I. Bang Andersen, M. B. Søndergaard Svendsen, A. L. Risgaard, C. Sander Danstrup, T. Todsén, M. G. Tolsgaard, and M. L. Friis, “Enabling micro-assessments of skills in the simulated setting using temporal artificial intelligence-models,” *Medical Teacher*, vol. 48, no. 3, pp. 415–424, 2026, doi: 10.1080/0142159X.2025.2555353.
- [15] S. Kulkarni, A. Saakyan, T. Chakrabarty, and S. Muresan, “A Report on the FigLang 2024 Shared Task on Multimodal Figurative Language,” In D. Ghosh, S. Muresan, A. Feldman, T. Chakrabarty, & E. Liu (Eds.), *FigLang 2024 - 4th Workshop on Figurative Language Processing, Proceedings of the Workshop* (pp. 115–119). Association for Computational Linguistics (ACL), 2024, doi: 10.18653/v1/2024.figlang-1.16.
- [16] M. Rahmanian, A. Sami, and Y. Yu, “Challenges and feasibility of multimodal LLMs in ER diagram evaluation,” *Cogent Education*, 12(1), 2025, doi: 10.1080/2331186X.2025.2590901.
- [17] C. J. Hellín, F. Calles-Esteban, A. Valledor, J. Gómez, S. Otón-Tortosa, and A. Tayebi, “Enhancing Student Motivation and Engagement through a Gamified Learning Environment,” *Sustainability (Switzerland)*, 15(19), 2023, doi: 10.3390/su151914119.
- [18] M. Liu and S. Liu, “AI in Digital Education: A Review,” *GAITDI 2025 - International Conference on Implementing Generative AI into Telecommunication and Digital Innovation 2025*, 284–288, 2026, doi: 10.1145/3771792.3773893.
- [19] R. S. Mansur, “Code, Test, Battle: Gamifying CS2 Through Adversarial Programming Tournaments,” *SIGCSE TS 2026 - Proceedings of the 57th ACM Technical Symposium on Computer Science Education V.2*, 1441–1442, 2026, doi: 10.1145/3770761.3777362.
- [20] J. Yoo, J. Park, M. Ha, and C. Mae Lagmay Darang, “Exploring Pre-Service Teachers’ Cognitive Processes and Calibration with an Unsupervised Learning-Based Automated Evaluation System,” *SAGE Open*, 14(3), 2024, doi: 10.1177/21582440241262864.
- [21] M. Yoshida, Y. Duangchinda, and N. Ruangrit, “AI-Supported Learning in Online Discussion Forums: A Scoping Review,” *Electronic Journal of E-Learning*, vol. 24, no. 1, pp. 19–35, 2026, doi: 10.34190/ejel.24.1.4409.
- [22] I. Choi, J. Hao, C. Li, M. Fauss, and J. Novák, “AutoESD: An Automated System for Detecting Nonauthentic Texts for High-Stakes Writing Tests,” *ETS Research Report Series*, vol. 2024, no. 1, pp. 1–16, 2024, doi: 10.1002/ets2.12383.
- [23] M. Li, “A study on the evaluation of the effect of the application of digital tools in project-based learning,” *Proceedings of 2025 International Conference on Digital Education and Information Technology, DEIT 2025*, 207–211, 2025, doi: 10.1145/3732299.3732337.
- [24] M. F. Astudillo, K. Krämer, and A. Arteaga, “Using dynamic life cycle assessment to evaluate the effects of industry digitalization: A steel case study,” *Journal of Industrial Ecology*, vol. 28, no. 4, pp. 942–952, 2024, doi: 10.1111/jiec.13510.
- [25] Z. Fan, W. Wang, X. Wu, and D. Zhang, “SedarEval: Automated Evaluation using Self-Adaptive Rubrics,” In Y. Al-Onaizan, M. Bansal, & C. Y.-N. (Eds.), *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024* (pp. 16916–16930). Association for Computational Linguistics (ACL), 2024, doi: 10.18653/v1/2024.findings-emnlp.984.
- [26] N. Gligoričević, S. D. Popović, V. Nikolić, D. Viduka, and S. Popović, “A Hybrid SWARA-NWA Framework for Evaluating AI-Based Image Recognition Algorithms in Educational Technology Applications,” *International Journal of Cognitive Research in Science, Engineering and Education*, vol. 13, no. 3, pp. 719–735, 2025, doi: 10.23947/2334-8496-2025-13-3-719-735.
- [27] A. Hambley, Y. Yesilada, M. Vigo, and S. Harper, “Optimising web accessibility evaluation: Population sourcing methods for web accessibility evaluation,” *International Journal of Human Computer Studies*, 198, 2025, doi: 10.1016/j.ijhcs.2025.103472.
- [28] R. Pires, R. Malaquias Junior, and R. Nogueira, “Automatic Legal Writing Evaluation of LLMs,” In J. Maranhao (Ed.), *20th International Conference on Artificial Intelligence and Law, ICAIL 2025 - Proceedings of the Conference* (pp. 420–424). Association for Computing Machinery, Inc, 2025, doi: 10.1145/3769126.3769227.

- [29] J. D. Quespaz-Sanchez and T. Calle-Jimenez, "A Closer Look at MOOC Accessibility: Comparing Local and Global Platforms with WCAG 3.0 in Mind," *International Journal of Information and Education Technology*, vol. 14, no. 7, pp. 1006–1015, 2024, doi: 10.18178/ijiet.2024.14.7.2128.
- [30] K. Xu, T. Yu, W. Hou, Y. Cheng, L. Li, X. Jiang, L. Shang, Q. Liu, and W. Li, "Learning to Align Multi-Faceted Evaluation: A Unified and Robust Framework," In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (pp. 9488–9502). Association for Computational Linguistics (ACL), 2025, doi: 10.18653/v1/2025.findings-acl.494.
- [31] M. Xue and D. Sun, "Design of online examination evaluation model for college students based on fuzzy reasoning system," *Discover Computing*, 28(1), 2025, doi: 10.1007/s10791-025-09855-0.
- [32] T. Zhang, L. Patterson, B. Webb, Z. Jin, and M. Beiting-Parrish, "Evaluating generative AI chatbots for large-scale assessment data: comparing LLM-as-a-judge and human ratings," *Large-Scale Assessments in Education*, 14(1), 2026, doi: 10.1186/s40536-026-00287-w.
- [33] Y. Zhao, W. Chen, Z. Xu, M. Patwardhan, Y. Liu, C. Wang, L. Vig, and A. Cohan, "ABGEN: Evaluating Large Language Models in Ablation Study Design and Evaluation for Scientific Research," In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 12479–12491). Association for Computational Linguistics (ACL), 2025, doi: 10.18653/v1/2025.acl-long.611.
- [34] H. Li, J. Chen, Q. Ai, Z. Chu, Y. Zhou, Q. Dong, and Y. Liu, "CalibraEval: Calibrating Prediction Distribution to Mitigate Selection Bias in LLMs-as-Judges," In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 16537–16552). Association for Computational Linguistics (ACL), 2025, doi: 10.18653/v1/2025.acl-long.808.
- [35] A. S. Panfilova and E. A. Valueva, "Deep Learning and Explainable AI for Creativity Scoring in TCT-DP Form B," *Psychology. Journal of Higher School of Economics*, vol. 22, no. 4, pp. 721–735, 2025, doi: 10.17323/1813-8918-2025-4-721-735.
- [36] J.-W. Chen, H.-L. Tu, C.-H. Chang, W.-C. Hsu, P.-C. Wang, C.-H. Liao, and M. Chen, "Automated Evaluation of Reflection and Feedback Quality in Workplace-Based Assessments by Using Natural Language Processing: Cross-Sectional Competency-Based Medical Education Study," *JMIR Medical Education*, vol. 11, e81718, Oct. 2025, doi: 10.2196/81718.
- [37] Y. V He, M. Chowdhury, J. Joselowitz, A. Higham, and E. Lim, "ASTRID - An Automated and Scalable TRIaD for the Evaluation of RAG-based Clinical Question Answering Systems," In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (pp. 16700–16716). Association for Computational Linguistics (ACL), 2025, doi: 10.18653/v1/2025.findings-acl.857.
- [38] M. Li, C. Li, Y. Chen, and G.-J. Hwang, "Effects of an automated evaluation mechanism on students' writing performance and higher-order thinking in an AR-based formative-peer-assessment learning mode," *Education and Information Technologies*, vol. 30, no. 15, pp. 21889–21928, 2025, doi: 10.1007/s10639-025-13611-8.
- [39] E. I. Shpit and P. M. McCarthy, "Novice Russian Research Writing: Prepositions and Prepositional Phrases," *Journal of Language and Education*, vol. 9, no. 2, pp. 146–159, 2023, doi: 10.17323/jle.2023.15897.
- [40] Y. Xu, J. Chen, and Y. Pan, "AI-Augmented Learning in Translation Education: An Analysis of Model Performance and Pedagogical Appropriateness Based on Automated Evaluation Metrics," *Proceedings of 2025 3rd International Conference on Information Education and Artificial Intelligence, ICIEAI 2025*, 1146–1151, 2026, doi: 10.1145/3799457.3799645.
- [41] T. Zhang, L. Patterson, B. Webb, Z. Jin, and M. Beiting-Parrish, "Evaluating generative AI chatbots for large-scale assessment data: comparing LLM-as-a-judge and human ratings," *Large-Scale Assessments in Education*, 14(1), 2026, doi: 10.1186/s40536-026-00287-w.
- [42] A. Nandy, Y. Agarwal, A. Patwa, M. M. Das, A. Bansal, A. Raj, P. Goyal, and N. Ganguly, "YesBut: A High-Quality Annotated Multimodal Dataset for evaluating Satire Comprehension capability of Vision-Language Models," In Y. Al-Onaizan, M. Bansal, & C. Y.-N. (Eds.), *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing. Proceedings of the Conference* (pp. 16878–16895). Association for Computational Linguistics (ACL), 2024, doi: 10.18653/v1/2024.emnlp-main.937.
- [43] H. Nguyen and S. Park, "Providing Automated Feedback on Formative Science Assessments: Uses of Multimodal Large Language Models," *15th International Conference on Learning Analytics and Knowledge, LAK 2025*, 803–809, 2025, doi: 10.1145/3706468.3706480.
- [44] Q. Wang and Y. Li, "Multi-Model Fusion for Street Visual Quality Evaluation," *ISPRS International Journal of Geo-Information*, 15(4), 2026, doi: 10.3390/ijgi15040158.
- [45] Y. Li, M. Raković, N. Srivastava, X. Li, Q. Guan, D. Gašević, and G. Chen, "Can AI support human grading? Examining machine attention and confidence in short answer scoring," *Computers & Education*, vol. 228, Art. no. 105244, Apr. 2025, doi: 10.1016/j.compedu.2025.105244.
- [46] H. Wu, Z. Shen, S. Hou, J. Liang, H. Jiao, Y. Qing, X. Zhang, X. Li, Z. Gui, X. Guan, and L. Xiang, "AutoGEEval: A Multimodal and Automated Evaluation Framework for Geospatial Code Generation on GEE with Large Language Models," *ISPRS International Journal of Geo-Information*, 14(7), 2025, doi: 10.3390/ijgi14070256.
- [47] D. Meiller, M. Wiehl, and L. Bäuml, "Automated Evaluation of Games programmed in Computer Science Assessments The use of CodeRunner to validate code written in the Processing Language," *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE*, vol. 2, pp. 775–776, 2024, doi: 10.1145/3649405.3659516.
- [48] H. Li, Y. Wu, V. Schlegel, R. Batista-Navarro, T. Madusanka, I. Zahid, J. Zeng, X. Wang, X. He, Y. Li, and G. Nenadic, "Which Side Are You On? A Multi-task Dataset for End-to-End Argument Summarisation and Evaluation," In K. L.-W., A. Martins, & V. Srikumar (Eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (pp. 133–150). Association for Computational Linguistics (ACL), 2024, doi: 10.18653/v1/2024.findings-acl.9.

- [49] J. Chen and J. Mueller, "Quantifying Uncertainty in Answers from any Language Model and Enhancing their Trustworthiness," In K. L.-W., A. F. T. Martins, & V. Srikumar (Eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 5186–5200). Association for Computational Linguistics (ACL), 2024, doi: 10.18653/v1/2024.acl-long.283.
- [50] Y. Zhu, Q. Chen, and M. Zhong, "Using Generative Artificial Intelligence to Evaluate the Quality of Chinese Environmental Information Disclosure in Chemical Firms," *Sustainability* (Switzerland), 17(24), 2025, doi: 10.3390/su172411348.
- [51] C. Anghel, A. A. Anghel, E. Pecheanu, A. Cocu, M. V Craciun, P. Iacobescu, A. S. Balau, and C. A. Andrei, "GraderAssist: A Graph-Based Multi-LLM Framework for Transparent and Reproducible Automated Evaluation," *Informatics*, 12(4), 2025, doi: 10.3390/informatics12040123.