International Journal of Learning, Teaching and Educational Research
Vol. 25, No. 7, pp. 689-715, July 2026
https://doi.org/10.26803/ijlter.25.7.31
Received May 1, 2026; Revised Jun 1, 2026; Accepted Jul 5, 2026

AI-driven Assessment in Education: A Bibliometric Analysis and Framework for Assessment Design and Governance

Shamsiah Banu Mohamad Hanefar*, **Maryam Ikram**
Faculty of Education and Liberal Arts
INTI International University, Malaysia

Mutia Sobihah Abd Halim
Faculty of Business and Management
Universiti Sultan Zainal Abidin, Malaysia

Farah Naaz Abd Yunos
Centre for English Language Studies
Sunway University, Malaysia

Anika Rahman
School of Social Sciences, Humanities and Language
Bangladesh Open University, Dhaka, Bangladesh

Abstract. The rapid advancement of artificial intelligence (AI), particularly generative AI and large language models, has transformed educational assessment practices, raising concerns regarding assessment validity, academic integrity, human judgement, and governance. This study aims to map the intellectual landscape of AI-driven assessment research and synthesise key developments that inform the design, implementation, and governance of AI-supported assessment. A bibliometric analysis was conducted on 820 Scopus-indexed publications published between 2015 and 2025. Using VOSviewer, publication trend analysis, keyword co-occurrence analysis, and bibliographic coupling analysis were employed to examine research productivity, thematic evolution, and intellectual structures within the field. The findings reveal a substantial increase in publications following the emergence of generative AI technologies, particularly between 2023 and 2025. Four major thematic areas were identified: generative AI and educational innovation, assessment and evaluation, AI technologies and learning systems, and disciplinary applications. Bibliographic coupling

Citation:
Hanefar, S, B, M., Ikram, M., Halim, M, S, A., Yunos, F, N, A., & Rahman, A. (2026). AI-driven Assessment in Education: A Bibliometric Analysis and Framework for Assessment Design and Governance. *International Journal of Learning, Teaching and Educational Research*, 25(7), 689–715. https://doi.org/10.26803/ijlter.25.7.31

*Corresponding author: *Shamsiah Banu Mohamad Hanefar, shamsiah.mhanefar@newinti.edu.my*

analyses further indicate growing intellectual convergence around assessment redesign, human-AI collaboration, academic integrity, and responsible AI use. These findings suggest that AI-driven assessment has evolved beyond technological implementation toward a broader socio-technical discourse encompassing pedagogical, ethical, and governance considerations. Based on the synthesis of the findings, this study proposes the Human-AI Governance and Design (HAGD) Framework, comprising three interdependent dimensions: Assessment Design, Human-AI Role Allocation, and Governance and Oversight. The framework provides a structured foundation for the responsible integration of AI into educational assessment while maintaining educational quality, fairness, transparency, and public trust. The study contributes to the growing discourse on AI-driven assessment and offers practical insights for educators, assessment designers, and policymakers.

Keywords: artificial intelligence; generative AI; educational assessment; bibliometric analysis; human-AI collaboration; assessment governance; education

1. Introduction

The convergence of advanced technologies and data-driven methodologies has fundamentally reshaped the modern educational landscape. Within this transformative environment, artificial intelligence (AI) has emerged as a particularly influential force, redefining traditional paradigms of teaching, learning, and, most notably, assessment practices (Cope et al., 2021; Mao et al., 2024). The progression of artificial intelligence in educational assessment has undergone significant technological and conceptual transformations, beginning with early explorations of computer-assisted instruction in the mid-20th century, which laid the foundation for integrating technology into pedagogical environments (Vashishth et al., 2024).

This historical context is critical for understanding the subsequent evolution toward more sophisticated systems. Initially, the primary AI applications in assessment were rule-based systems, often referred to as Automatic Assessment Tools (AATs). These tools were highly effective at evaluating objective, well-structured tasks, such as multiple-choice questions, short-answer responses, and programming assignments, by employing static analysis techniques, including syntax checking, and dynamic analysis via unit testing (Gao, 2025). These early systems were characterised by a focus on automating the grading of structured data, thereby serving a clear, albeit limited, function.

A significant and far-reaching paradigm shift has occurred with the advent of generative artificial intelligence (GenAI) and large language models (LLMs) (Mao et al., 2024; Xia et al., 2024). Unlike AATs, which operate according to predefined rules, LLMs are trained on large-scale datasets and use natural language processing (NLP) to evaluate complex, open-ended tasks, such as essays and discussions (Fagbohun et al., 2024; Xu et al., 2024). This technological leap enables the assessment of subjective work, fundamentally altering what can be automatically graded. The generative capacity of these models, which can create

human-like text, is the very attribute that enables their use in evaluating complex submissions while also introducing new challenges related to academic integrity (Giannakos et al., 2025).

Research has systematically benchmarked the performance of these tools, quantifying their capabilities and limitations. A multi-institutional study of engineering education assessments found that, as of early 2023, ChatGPT-3.5 could achieve passable grades on a wide range of assessments, demonstrating particular strengths in written activities (Nikolic et al., 2023). However, it often struggled with numerical tasks and assessments requiring interpretation of diagrams or tables (Nikolic et al., 2023). An updated study conducted a year later revealed that newer tools, including ChatGPT-4, Copilot, and Gemini, had made substantial performance gains (Nikolic et al., 2024). This updated analysis found that these more advanced tools demonstrated promising, albeit imperfect, capabilities with complex images and documents, further blurring the lines between what is considered a secure or bulletproof assessment (Nikolic et al., 2024).

This progression from rule-based automated assessment tools (AATs) to sophisticated generative large language models (LLMs) is not merely a technological upgrade; it represents a profound shift in purpose. Early tools were primarily designed to automate the evaluation of simple outputs of the learning process, such as quizzes or structured assignments. The capabilities of modern GenAI models, by contrast, move the technology from a tool of simple efficiency to a potential partner in complex pedagogical processes, including brainstorming, outlining, and drafting (Luckin et al., 2016; Holmes et al., 2019). This presents a new layer of challenges. If a system can produce a passable essay or project component, it can undermine the very purpose of assessments designed to measure a student's own cognitive process and intellectual effort (Sweeney, 2023). Some benchmarked studies highlight a critical concern: the capabilities of these systems are evolving more rapidly than institutions can redesign assessment practices or establish policies to safeguard academic integrity (Zawacki-Richter et al., 2019; Chen et al., 2020).

Given this rapid development, there is an urgent need to systematically study AI for assessment to understand both its current applications and future direction. While scattered studies exist, the field lacks a comprehensive mapping of research trends, influential works, and thematic clusters that indicate the direction of scholarship. For instance, a search of related documents on bibliometric studies in the Scopus database yielded only five results, of which two are unrelated to education; the remaining three are closely related, but one is a conference paper, and two are lecture notes.

The current study is significant because it encompasses a broader range of bibliometric analyses, including coherence analysis and bibliographic coupling, thereby providing a more detailed understanding of research trends in the field. Bibliometric methods enable researchers to uncover knowledge structures, identify highly cited works shaping the discourse, and track emerging themes

such as adaptive feedback, fairness, ethical implications, and policy frameworks. By using bibliometric techniques, scholars can develop a clearer picture of the past, present, and potential future of AI in assessment, helping educational systems to make informed and responsible decisions about integrating these technologies (Prasad et al., 2025).

While existing reviews and bibliometric studies have mapped the growth of artificial intelligence research in education, fewer studies explicitly translate these patterns into guidance for assessment design, validity, and governance. Responding to this gap, the present study not only maps the intellectual structure of AI-driven assessment research but also synthesises its implications into a conceptual framework intended to inform assessment design decisions, human–AI role allocation, and institutional policy in AI-rich educational environments.

Specifically, the research questions (RQ) this study aims to answer are:
RQ1: What are the publication trends and research patterns in AI-driven assessment in education between 2015 and 2025?
RQ2: What thematic clusters and intellectual structures characterise AI-driven assessment research?
RQ3: How can the bibliometric findings inform the development of a framework for AI-driven assessment design and governance?

2. Methodology

This study employed a systematic bibliometric approach to analyse research trends related to AI-driven assessment in education. The methodology was informed by the principles of Systematic Literature Review (SLR), which provides a rigorous and transparent process for identifying, screening, and synthesising scholarly literature (Xiao & Watson, 2019). Integrating bibliometric analysis with systematic review principles enables a structured examination of the development, thematic evolution, and intellectual structure of research within the field (Table 1, Figure 1).

The Scopus database was selected as the primary data source due to its extensive and reliable coverage of peer-reviewed journals, conference proceedings, books, and interdisciplinary scholarly publications (Harzing & Alakangas, 2016). To ensure conceptual precision and thematic relevance, the search strategy was restricted to the TITLE field. Preliminary, broader searches using TITLE-ABS-KEY yielded an excessively large and heterogeneous dataset, including numerous publications in which AI or assessment appeared only marginally. Therefore, the title-focused strategy was adopted to prioritise publications explicitly centred on AI-driven assessment in educational contexts.

The search string combined AI-related terms such as “artificial intelligence,” “AI,” “generative AI,” “GenAI,” “ChatGPT,” “large language model*,” and “LLM*” with assessment-related terms including “assessment,” “evaluation,” “grading,” “feedback,” “e-assessment,” “automated assessment,” and “educational assessment,” together with educational context terms such as “education,” “higher education,” “university,” and “college.” The initial search without

restrictions yielded 1,115 records. Subsequently, publication year and language limits were applied, restricting the dataset to English-language publications from 2015 to 2025, resulting in a final dataset of 820 records.

The retrieved records underwent data cleaning to ensure consistency and accuracy prior to analysis. This process included verification of document titles, authors' names, and Digital Object Identifiers (DOIs), as well as the removal of incomplete or irrelevant records identified during screening. Detailed information regarding the search strategy, inclusion and exclusion criteria, and screening procedure is presented in Table 1 and Figure 1.

The bibliometric analysis was conducted using VOSviewer software (van Eck & Waltman, 2010), which enables the construction and visualisation of bibliometric networks. The default VOSviewer clustering parameters were applied, with a resolution parameter of 1.00 and a maximum of 1000 iterations. Co-occurrence analysis was employed to identify major keyword clusters and thematic relationships within the literature, while bibliographic coupling analysis was used to examine intellectual linkages and influential research contributions across the field. In addition to descriptive bibliometric mapping, the findings were interpretively synthesised to derive implications related to assessment design, governance, ethics, validity, fairness, and the evolving role of AI in educational assessment practices.

By integrating systematic review principles with bibliometric techniques, the methodology provides a robust and transparent foundation for examining how AI-driven assessment research has evolved over time and how emerging trends may inform future assessment design, human–AI role allocation, and governance considerations in educational settings.

Table 1: Search string

Database	**Scopus**
Time Period	2015-2025
Search field	Article title
Search keywords	Round 1: TITLE ((“artificial intelligence" OR AI OR "generative AI" OR GenAI OR ChatGPT OR "large language model*" OR LLM*) AND (assessment OR evaluation OR grading OR feedback OR "e-assessment" OR "automated assessment" OR "educational assessment") AND (education OR "higher education" OR university OR universities OR college OR colleges)) (1,115 documents) Round 2: TITLE ((“artificial intelligence" OR AI OR "generative AI" OR GenAI OR ChatGPT OR "large language model*" OR LLM*) AND (assessment OR evaluation OR grading OR feedback OR "e-assessment" OR "automated assessment" OR "educational assessment") AND (education OR "higher education" OR university OR universities OR college OR colleges)) AND LIMIT-TO (LANGUAGE, "English")) AND PUBYEAR > 2014 AND PUBYEAR < 2026 (820 documents)

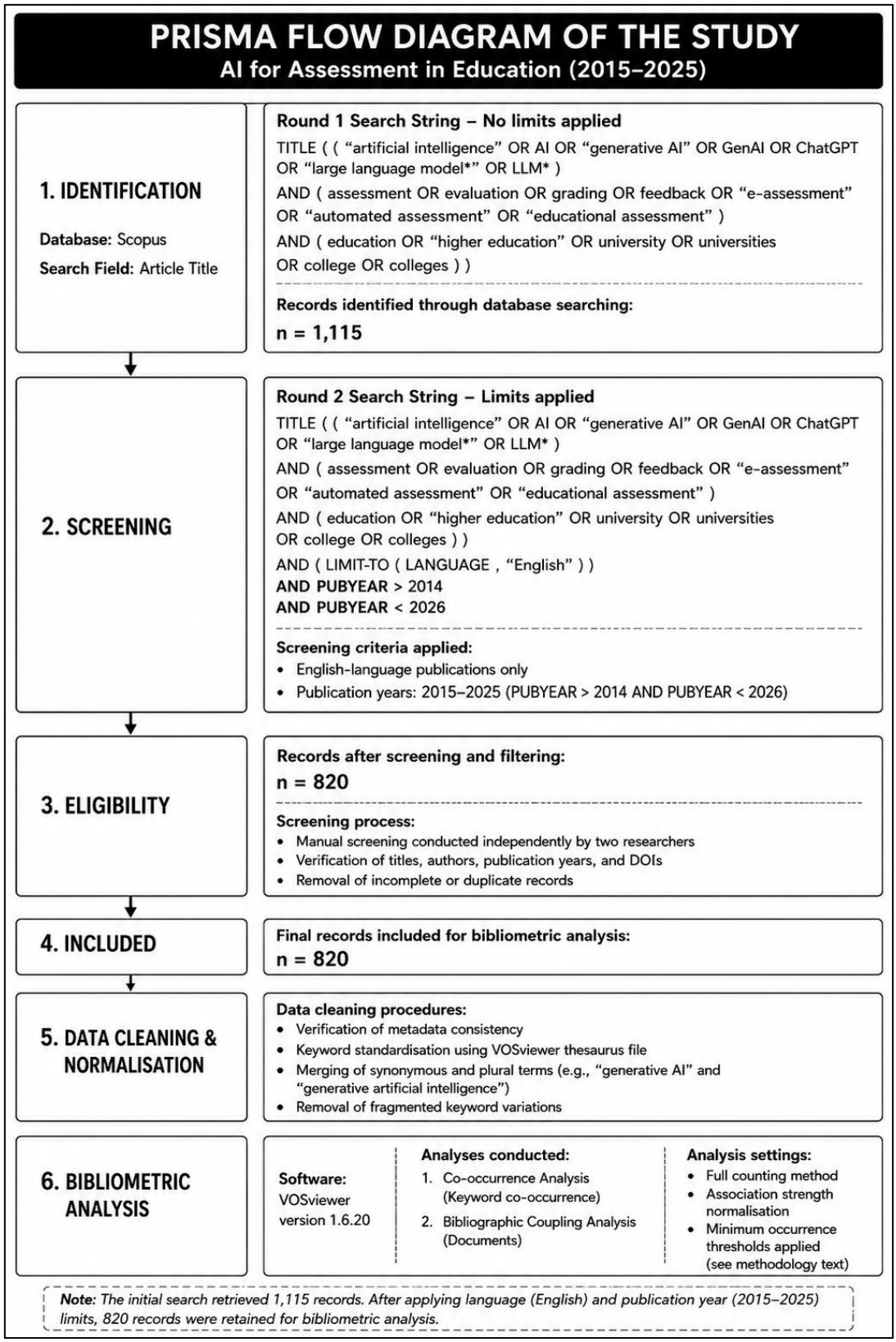

Figure 1: PRISMA flow diagram of the study
(Adapted from Mohamad Hanefar et al., 2025)

2.1 Data Cleaning, Screening, and Bibliometric Analysis Procedures

The retrieved records were exported from the Scopus database in CSV format and processed using Microsoft Excel and VOSviewer version 1.6.20 for data cleaning, screening, and bibliometric analysis. Prior to analysis, the dataset underwent a systematic verification process to ensure accuracy, consistency, and relevance. This process included checking document titles, author names, publication years, source titles, and Digital Object Identifiers (DOIs). Duplicate and incomplete records were manually reviewed and removed where necessary.

The screening process was conducted manually by two researchers independently to improve consistency and minimise selection bias. The researchers reviewed the retrieved records against the predefined inclusion and exclusion criteria, focusing specifically on publication year and language. Any discrepancies identified during the screening process were discussed and resolved through consensus. The final screening process retained 820 English-language publications from 2015 to 2025 for bibliometric analysis.

To improve the reliability and interpretability of the bibliometric mapping, keyword normalisation procedures were applied before conducting the co-occurrence analysis. Variations in terminology, plural forms, abbreviations, and synonymous keywords were standardised using a VOSviewer thesaurus file. For example, terms such as "generative artificial intelligence" and "generative AI" were merged into a single standardised keyword, while singular and plural variations such as "large language model" and "large language models" were unified. This process reduced fragmentation within the keyword network and improved the accuracy of thematic clustering.

The bibliometric analysis was conducted using VOSviewer version 1.6.20 (van Eck & Waltman, 2010). The software was selected for its ability to effectively construct, visualise, and analyse bibliometric networks. Co-occurrence analysis was employed to identify major research themes and keyword clusters within the field, while bibliographic coupling analysis was used to examine intellectual relationships and influential contributions among publications. The analysis adopted the full counting method and association strength normalisation to measure relationships between nodes within the bibliometric networks. Minimum occurrence thresholds were applied to improve network clarity and reduce noise from low-frequency terms.

The generated bibliometric maps were interpreted based on node size, link strength, clustering patterns, and thematic relationships. Larger nodes represented higher keyword frequency or publication influence, while stronger linkages indicated closer thematic or intellectual relationships between items. The findings from these analyses were subsequently synthesised to support discussion on emerging trends, assessment design implications, governance considerations, and the evolving role of AI in educational assessment practices.

2.2 Data Analysis Steps

2.2.1 Step 1: Research trends and patterns

This step provides a descriptive analysis of the collected documents to illustrate the field's growth over time and the dominant publication formats. By examining annual outputs, the analysis highlights the increasing scholarly attention to AI in assessment, particularly following advances in machine learning and the development of adaptive testing. Document type analysis (e.g., journal articles, conference papers, policy papers) further reveals the primary dissemination channels and the relative emphasis placed on different forms of scholarly communication within the field.

2.2.2 Step 2: Co-occurrence analysis

Co-occurrence analysis is applied to uncover the key research themes in the field. By analysing how often keywords co-occur within the same documents, the method identifies central clusters that represent distinct themes (Klarin, 2024). These clusters capture both the technological and pedagogical dimensions of AI-driven assessment, as well as recurring concerns about fairness, bias, and transparency, thereby revealing the intellectual structure of the research landscape.

2.2.3 Step 3: Bibliographic coupling

The final step identifies influential contributions and intellectual linkages within the field. Bibliographic coupling links documents that share references, indicating common theoretical or conceptual foundations (Liu, 2017). When applied to authors and countries, this analysis reveals leading contributors and regional research concentrations, offering insight into how scholarly influence and thematic alignment are distributed across disciplines and geographic contexts.

3. Results and Discussions

3.1 Trends in Publication – Documents by Year

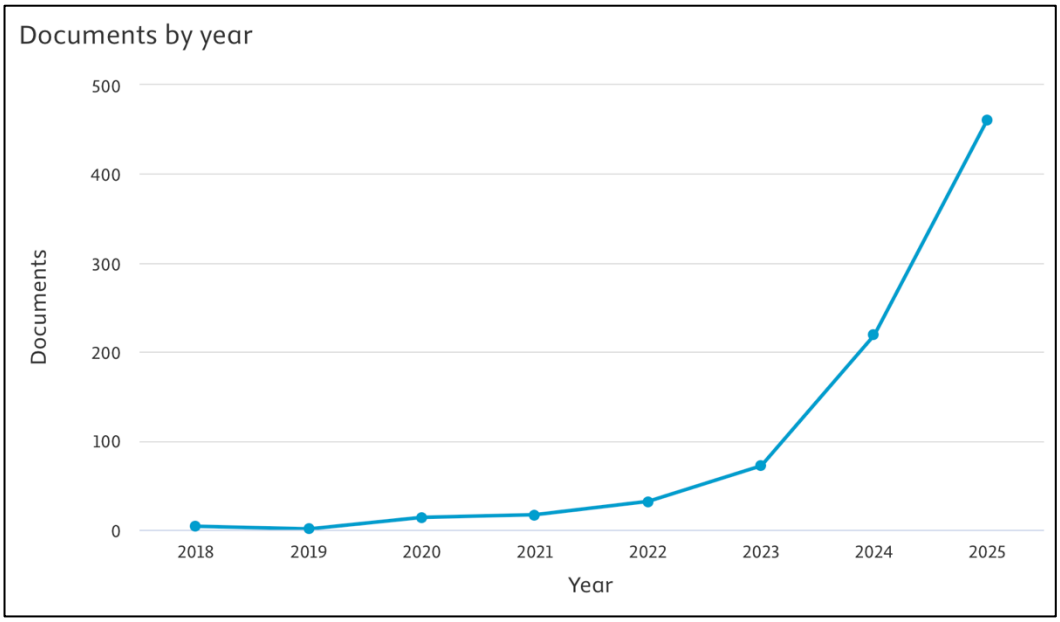

Figure 2: Documents by year

Based on the Scopus search (Figure 2), no publications related to AI for assessment in education were identified in the Scopus database between 2015 and 2017, indicating that the topic had not yet emerged as a significant area of scholarly attention during this period. Initial publications only began appearing in 2018 and remained very limited until 2021, with only a small number of documents produced annually. This reflects the early exploratory stage of research on AI-driven assessment practices in education.

A gradual increase in publication activity began in 2020, followed by a sharper rise in 2023, signalling growing academic interest in integrating AI into assessment and evaluation processes. The most significant growth occurred from 2023 onwards. In 2024, the number of publications increased sharply to more than 200, demonstrating that AI assessment had become a major research focus across educational disciplines. By 2025, publication output rose even more dramatically, reaching approximately 460 documents. This trend indicates that the field is still experiencing rapid expansion rather than entering a stabilisation phase.

This substantial increase can largely be attributed to the widespread adoption of generative AI technologies such as ChatGPT (Ray, 2023) and Gemini (Imran & Almusharraf, 2024), which intensified global discussions surrounding assessment design, academic integrity, automated feedback, personalised assessment, and the evolving role of educators in AI-supported learning environments. Consequently, recent studies have increasingly shifted from conceptual discussions toward empirical investigations and action research examining how assessment practices can be redesigned in response to AI-generated outputs.

For instance, several studies have explored the use of AI for personalised and low-stakes assessments (Halkiopoulos & Gkintoni, 2024; Vashishth et al., 2024). Researchers have also examined AI-driven formative assessments that provide students with immediate feedback (Zhai & Nehm, 2023; Vittorini et al., 2021). Some studies describe pilot implementations in which AI chatbots supported students in writing and problem-solving activities (Hamid et al., 2023), while others measured student engagement and learning outcomes associated with these technologies (Roca et al., 2024). These empirical findings suggest that AI can function as a continuous learning assistant, enabling students to receive rapid feedback and repeated practice opportunities without delays.

Another major research direction focuses on developing AI-resistant or AI-proof summative assessments. Instead of relying solely on traditional written assignments, educators have experimented with oral examinations, project-based assessments, authentic tasks, and context-specific assignments that require personal reflection or real-world application, rather than AI-generated responses (Awidi, 2024; Vittorini et al., 2021). These studies provide practical evidence regarding the implementation challenges, effectiveness, and implications of alternative assessment strategies in maintaining academic integrity.

In addition, some studies have investigated hybrid assessment models involving collaboration between AI systems and human evaluators. For example,

researchers have compared grading outcomes from AI tools such as ChatGPT with those from human instructors on the same assessment tasks (Jukiewicz, 2024; Villegas-Ch et al., 2024). Findings generally indicate that while AI can provide efficient, consistent scoring for lower-order tasks, human evaluators remain essential for assessing higher-order thinking, contextual understanding, detailed feedback, and ethical considerations (Molenaar, 2024; Tossell et al., 2024). This has led to growing support for blended assessment approaches in which AI assists with routine grading while educators focus on mentorship and deeper pedagogical engagement.

Collectively, these developments contributed directly to the sharp surge in publications observed from 2023 to 2025. The rapid expansion reflects not only growing scholarly interest but also the urgent need for educational institutions worldwide to address the disruptive impact of generative AI on assessment practices. Furthermore, the literature increasingly demonstrates a shift from exploratory discussions to evidence-based investigations and practical interventions, suggesting that AI-driven assessment research is entering a formative, highly active stage of development.

3.2 Co-occurrence Analysis

In this co-occurrence analysis, a total of 1,844 keywords were initially identified; however, only those that appeared at least twice were considered. This minimum occurrence threshold reduced the number of keywords to be analysed to 302. For each of these 302 keywords, the total strength of its connections with other co-occurring keywords was calculated. The highest-ranking keywords are those with the strongest total link strength and were then selected for further study.

Table 2: Top twenty keywords with the highest total link strength

No.	Keyword	Occurrences	Total Link Strength
1	Artificial intelligence	259	592
2	ChatGPT	112	336
3	Higher education	91	264
4	Generative AI	65	176
5	Assessment	58	170
6	Large Language Models	47	153
7	Education	47	121
8	Medical Education	27	102
9	Natural Language Processing	20	101
10	AI	28	90
11	Feedback	28	87
12	Generative Artificial Intelligence	34	84
13	Educational technology	24	82
14	Machine Learning	23	81
15	Large Language Model	19	67
16	Academic Integrity	23	65
17	Chatbot	14	60
18	Artificial Intelligence	25	58
19	NLP	11	55
20	Education Technology	6	40

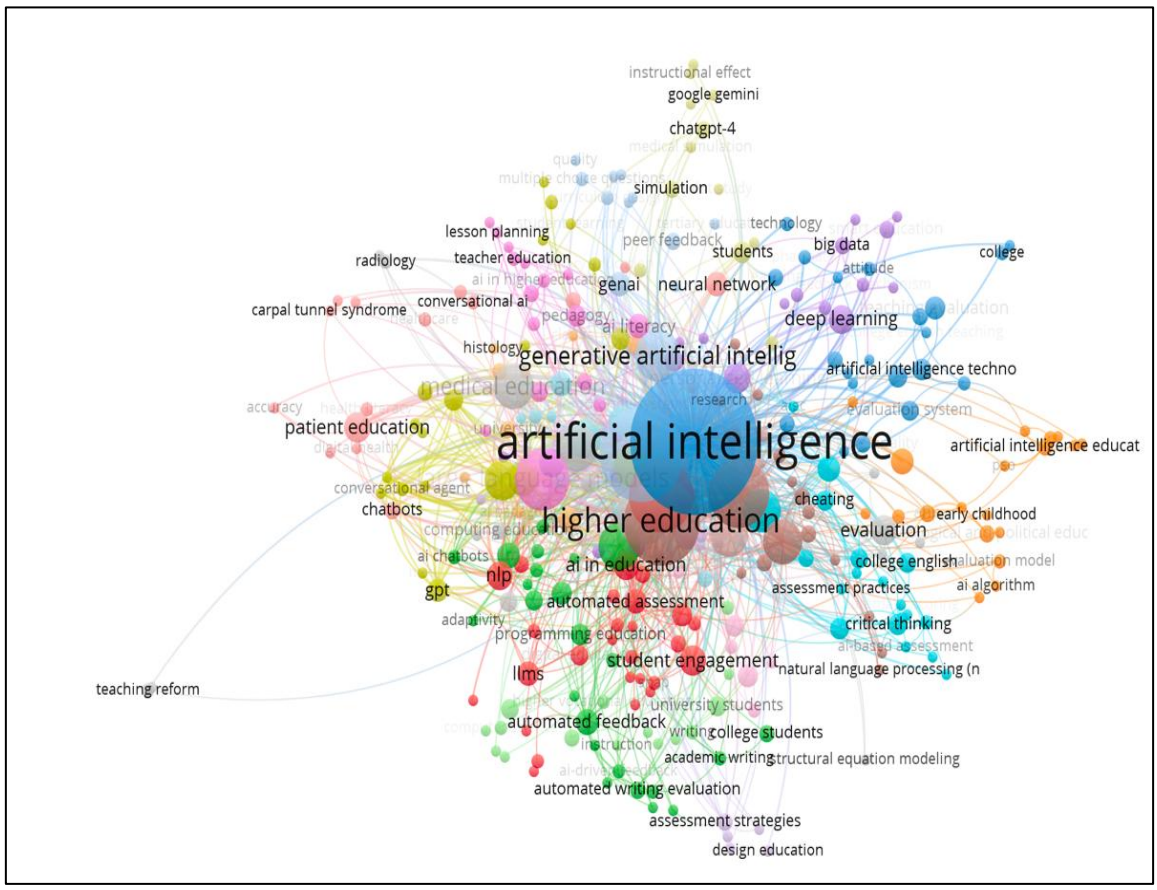

Figure 3: Co-occurrence analysis

The co-occurrence analysis (Table 2, Figure 3) reveals a highly interconnected research landscape centred on artificial intelligence (AI), which emerges as the dominant keyword and largest node in the network. Its strong connections with higher education, generative artificial intelligence, evaluation and assessment practices, students, and automated assessment indicate that AI-driven assessment research is predominantly situated within higher education and is increasingly influenced by advances in generative AI technologies. The prominence of these keywords suggests that research has evolved beyond examining AI as a general educational technology and now focuses on its implications for assessment, feedback, academic integrity, and student learning (Naidu & Sevnarayan, 2023; Smolansky et al., 2023).

The network visualisation reveals four major thematic clusters. The first cluster, Generative AI and Educational Innovation, centres on keywords such as generative artificial intelligence, GenAI, ChatGPT-4, Google Gemini, AI literacy, peer feedback, simulation, and lesson planning. This cluster reflects the growing influence of large language models on teaching, learning, and assessment. Studies in this area explore how generative AI supports content generation, personalised learning, feedback provision, and instructional design, while also raising concerns about ethical use, academic integrity, and AI literacy among educators and students (Naidu & Sevnarayan, 2023; Nikolic et al., 2024).

The second cluster, Assessment and Evaluation, includes keywords such as evaluation, assessment practices, automated assessment, automated feedback, automated writing evaluation, student engagement, critical thinking, and academic writing. This cluster directly reflects the focus of the present study and highlights growing interest in AI-assisted assessment systems. The strong interconnections among these keywords indicate that researchers are increasingly investigating how AI can support formative assessment, provide immediate feedback, enhance student engagement, and facilitate the evaluation of complex learning outcomes (Zhai & Nehm, 2023). At the same time, the coexistence of automated assessment and critical thinking suggests ongoing efforts to ensure that AI-supported assessment can evaluate higher-order cognitive skills rather than merely automate grading processes (Molenaar, 2024).

The third cluster, Artificial Intelligence Technology, comprises keywords such as deep learning, neural network, natural language processing, big data, and AI algorithm. These terms represent the technological foundations underpinning AI-driven assessment systems. Research within this area focuses on machine learning algorithms, predictive analytics, natural language processing techniques, and intelligent educational systems that support automated evaluation, feedback generation, and educational decision-making (Hooda et al., 2022; Zhai & Nehm, 2023). The prominence of these concepts demonstrates the continued importance of data-driven approaches in advancing AI-supported assessment practices. Furthermore, the close connectivity between technological and pedagogical keywords highlights the increasingly interdisciplinary nature of the field, where developments in artificial intelligence directly influence educational innovation and assessment design (Villegas-Ch et al., 2024; Xu et al., 2024).

The fourth cluster, Disciplinary Applications, encompasses keywords such as medical education, patient education, radiology, teacher education, programming education, and college English. This cluster demonstrates the widespread adoption of AI-assisted assessment across diverse educational disciplines. Medical education appears particularly prominent, reflecting AI's suitability for competency-based assessment, simulation training, and performance evaluation (Awidi, 2024). Beyond healthcare, AI-assisted assessment is increasingly applied in language education, teacher education, and programming education, supporting automated feedback, personalised learning, and skills assessment (Halkiopoulos & Gkintoni, 2024; Nikolic et al., 2024). The presence of these disciplinary keywords suggests that AI-driven assessment has evolved into a cross-disciplinary educational innovation, adaptable to diverse learning contexts and educational needs (Naidu & Sevnarayan, 2023; Smolansky et al., 2023).

A notable finding is the central position of generative artificial intelligence, which appears to be one of the most influential nodes in the AI network. The prominence of keywords such as ChatGPT-4, Google Gemini, AI literacy, and GenAI indicates a paradigm shift from earlier machine-learning-focused research toward generative AI-driven assessment innovation. Furthermore, the strong presence of terms such as cheating, assessment practices, critical thinking,

and evaluation suggests growing scholarly attention to academic integrity and assessment redesign in response to AI-generated content (Nikolic et al., 2024). Overall, the findings demonstrate that AI-driven assessment research has become increasingly interdisciplinary, integrating technological innovation with pedagogical and assessment-related concerns while focusing on the responsible and effective use of AI in educational evaluation.

3.3 Bibliographic Coupling

The bibliographic coupling analysis illustrates the intellectual structure of research on AI-driven assessment in education by mapping documents that share common references. Each node represents a publication, with its size reflecting coupling strength, and the connections indicating the degree of overlap in cited sources. The visualisation reveals several distinct clusters, suggesting that scholarship in this area is currently fragmented into sub-communities with varied thematic orientations.

3.3.1 By document

In this bibliographic coupling by document, documents with at least two citations are selected. Out of the 820 documents initially identified, 452 met this minimum threshold. For each of these 452 documents, the total strength of its bibliographic links to other documents was calculated. The documents with the greatest total link strength were then selected for further study.

Table 3: Top ten bibliographic coupling by document

No.	Document	Citations	Total Link Strengths
1	Lelescu (2024)	5	141
2	Flodén (2025)	54	140
3	Coskun (2024)	25	130
4	Agostini (2024b)	19	126
5	Cildir (2024)	6	113
6	Oc (2025)	35	109
7	Khlaif (2024)	121	108
8	Mustapha (2024)	21	105
9	Gamage (2023)	68	105
10	Rudolph (2023)	1447	104

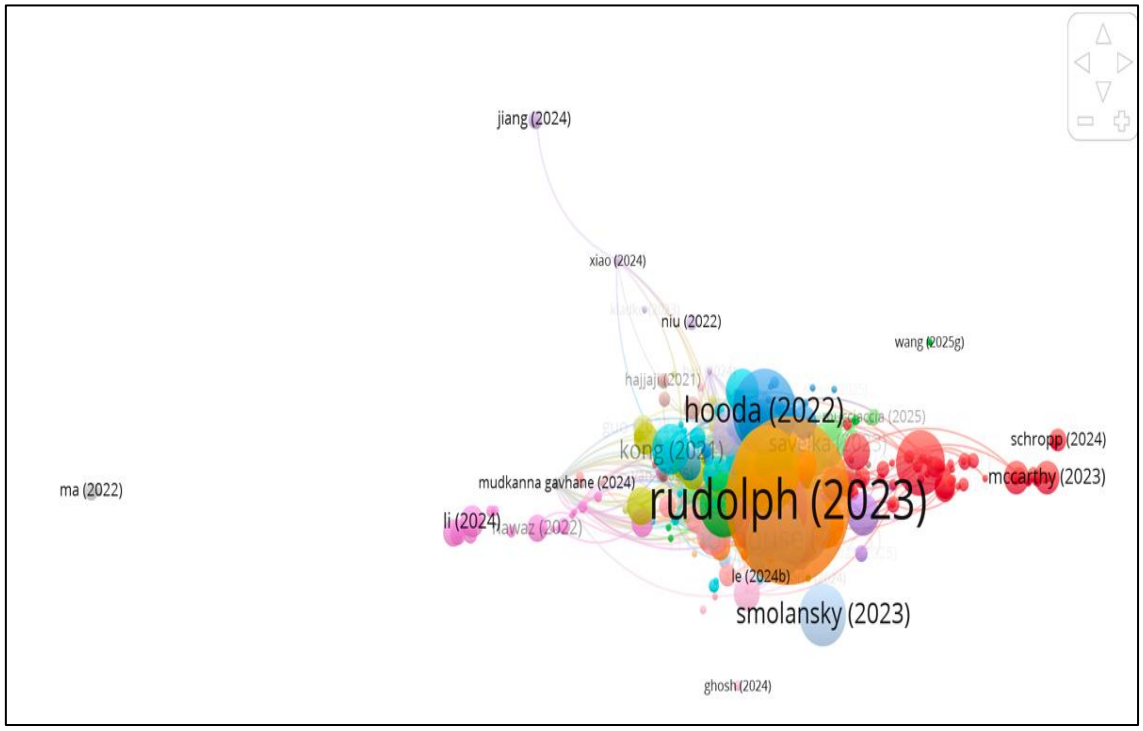

Figure 4: Bibliographic coupling by document

The bibliographic coupling analysis (Table 3, Figure 4) reveals the intellectual structure of research on AI-driven assessment in education by identifying documents that share similar reference bases. Among the top ten coupled documents, Lelescu and Kabiraj (2024) recorded the highest total link strength (TLS = 141), closely followed by Flodén (2025) (TLS = 140). Although several of these publications have accumulated relatively modest citation counts, their high coupling strengths indicate that they are strongly connected to the field's core knowledge base and share substantial intellectual foundations with other influential studies.

A notable finding is the presence of both highly cited and recently published works among the most strongly coupled documents. For example, Khlaif et al. (2024) received 121 citations while maintaining a strong coupling strength (TLS = 108), suggesting that it is both an influential and highly integrated publication in the literature. Similarly, Flodén (2025) (54 citations, TLS = 140) and Oc et al. (2025) (35 citations, TLS = 109) demonstrate that recent studies are already becoming central contributors to the evolving research landscape. This pattern indicates that the field is expanding rapidly, with newly published works quickly connecting to established streams of scholarship.

The network visualisation further reveals a densely connected core dominated by Rudolph (2023), which appears as the largest node despite ranking tenth in coupling strength (TLS = 104). The prominence of Rudolph (2023), together with influential studies such as Smolansky (2023) and Hooda (2022), suggests that these publications offer foundational perspectives widely referenced across multiple research themes. Their central positions indicate that they serve as

intellectual hubs linking diverse strands of research across AI, generative AI, assessment, and higher education.

Unlike a fragmented or linear network structure, the coupling map exhibits a relatively cohesive, interconnected knowledge base. Most highly coupled documents are clustered around a central core, indicating substantial overlap in the references used by researchers. This suggests that the field is gradually moving beyond its exploratory stage and developing a more consolidated body of knowledge. The close proximity among documents published between 2023 and 2025 further reflects the rapid convergence of research interests surrounding generative AI, assessment redesign, academic integrity, and AI-supported learning.

Overall, the bibliographic coupling analysis highlights a field that is simultaneously grounded in influential foundational studies and shaped by emerging contributions. The strong coupling strengths of recent publications such as Flodén (2025), Oc et al. (2025), and Lelescu and Kabiraj (2024) indicate that new research is building upon a shared intellectual foundation rather than developing in isolation. At the same time, the central roles of Rudolph (2023), Smolansky (2023), Hooda (2022), and Khlaif et al. (2024) demonstrate their importance in shaping the contemporary discourse on AI-driven assessment in education. Collectively, these findings suggest that the field is becoming increasingly mature, with a growing level of intellectual integration and convergence around key research themes.

3.3.2 By authors

In the bibliographic coupling analysis by authors, the minimum number of documents per author was set to 2, and the minimum number of citations per author to 2. Out of a total of 2792 authors, 81 met these thresholds. For each of these 81 authors, the total number of bibliographic links with other authors was calculated, and those with the highest total link strength were identified as the most influential within the network.

Table 4: Top ten authors with the highest total link strength.

No.	Author	Documents	Citations	Total Link Strength
1	Belkina, Marina	2	375	805
2	Daniel, Scott	2	375	805
3	Grundy, Sarah	2	375	805
4	Haque, Rezwanul	2	375	805
5	Hassan, Ghulam M.	2	375	805
6	Lyden, Sarah	2	375	805
7	Neal, Peter	2	375	805
8	Nikolic, Sasha	2	375	805
9	Agostini, Daniele	2	25	443
10	Picasso, Federica	2	25	443

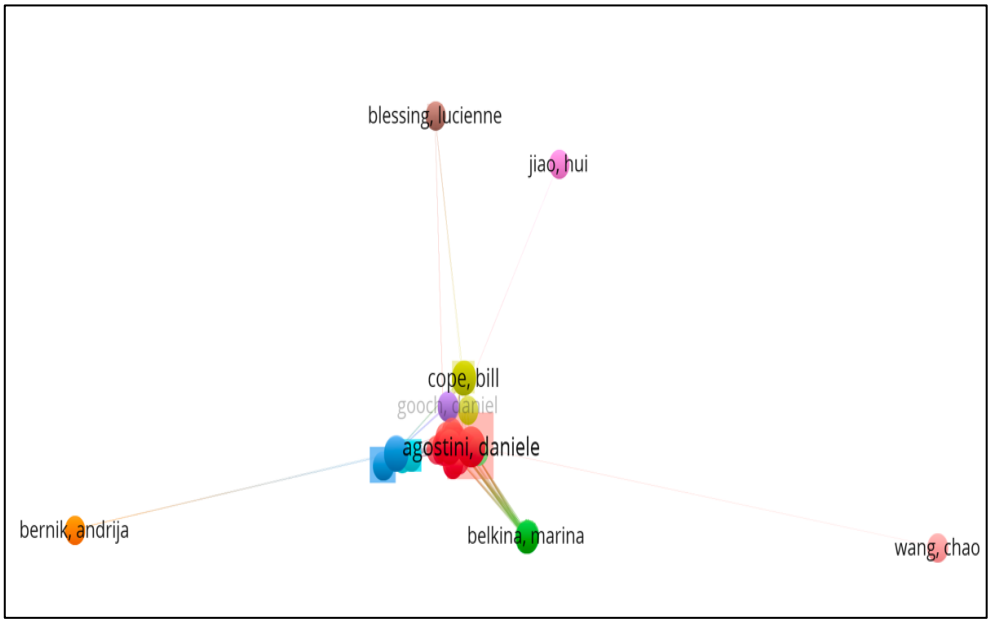

Figure 5: Bibliographic coupling by author

The bibliographic coupling analysis by author (Table 4, Figure 5) reveals a concentrated intellectual structure within the AI-driven assessment literature. A dominant group comprising Belkina, Marina; Daniel, Scott; Grundy, Sarah; Haque, Rezwanul; Hassan, Ghulam M.; Lyden, Sarah; Neal, Peter; and Nikolic, Sasha recorded identical bibliometric profiles, each contributing two documents, receiving 375 citations, and achieving the highest total link strength (TLS = 805).

Such uniformity suggests that these authors belong to a closely connected research stream and rely heavily on a shared body of references. According to Jarneving (2007), strong bibliographic coupling reflects common intellectual foundations and often indicates the presence of a specialised research front. In this case, the tightly connected cluster appears to be associated with emerging discussions on generative AI, assessment redesign, and higher education (Nikolic et al., 2024).

The network visualisation further demonstrates that these authors occupy the central position within the coupling map, reflecting their substantial influence on the field's current knowledge structure. This finding aligns with influential works by Rudolph et al. (2023), Smolansky et al. (2023), and Hooda et al. (2022), which have significantly shaped contemporary discussions on AI integration, educational assessment, and the implications of generative AI in higher education. The high concentration of coupling relationships among a limited number of authors suggests that the field is increasingly converging on a shared reference base.

A second cluster represented by Agostini and Picasso exhibits lower citation counts (25 citations each) but relatively high coupling strengths (TLS = 443). This pattern indicates that recent publications can become strongly embedded within the field despite not yet accumulating substantial citation impact. Their high

coupling strength suggests close alignment with prevailing research themes rather than the development of entirely distinct perspectives.

At the same time, peripheral authors such as Jiao, Blessing-Lucienne, Bernik, and Wang occupy more isolated positions within the network. Following Gazni and Didegah (2016), such authors may represent emerging or interdisciplinary areas that have yet to become fully integrated into the dominant research stream. While the central cluster promotes intellectual cohesion, the presence of peripheral contributors introduces diversity and potential avenues for future development.

Overall, the coupling structure reflects a field that is becoming increasingly consolidated around a common set of foundational studies. Consistent with Boyack and Klavans (2010), the presence of a strong central cluster suggests the emergence of a mature research front. However, as Glänzel and Czerwon (1995) argued, continued growth depends on balancing intellectual cohesion with the integration of new perspectives. Future research should therefore expand beyond existing reference patterns to encourage greater theoretical and methodological diversity within AI-driven assessment research.

3.3.3 By country

Based on the bibliographic coupling analysis by country, the minimum thresholds were set at 2 documents and 2 citations per country. Out of the 116 countries represented in the dataset, 71 met these thresholds. For each of these 71 countries, bibliographic coupling links to other countries were calculated to assess the strength of their research connections. The countries with the highest total link strength were then identified, reflecting their stronger collaborative and thematic alignment in the literature.

Table 5: Top ten countries with the highest total link strength

No.	Country	Documents	Citations	Total Link Strength
1	China	212	2095	6176
2	United States	106	3889	6016
3	United Kingdom	63	1021	5519
4	Malaysia	23	159	2937
5	Australia	31	939	2793
6	Hong Kong	21	1118	2487
7	India	58	691	2323
8	Germany	31	361	2214
9	Saudi Arabia	22	612	2208
10	Türkiye	29	384	2164

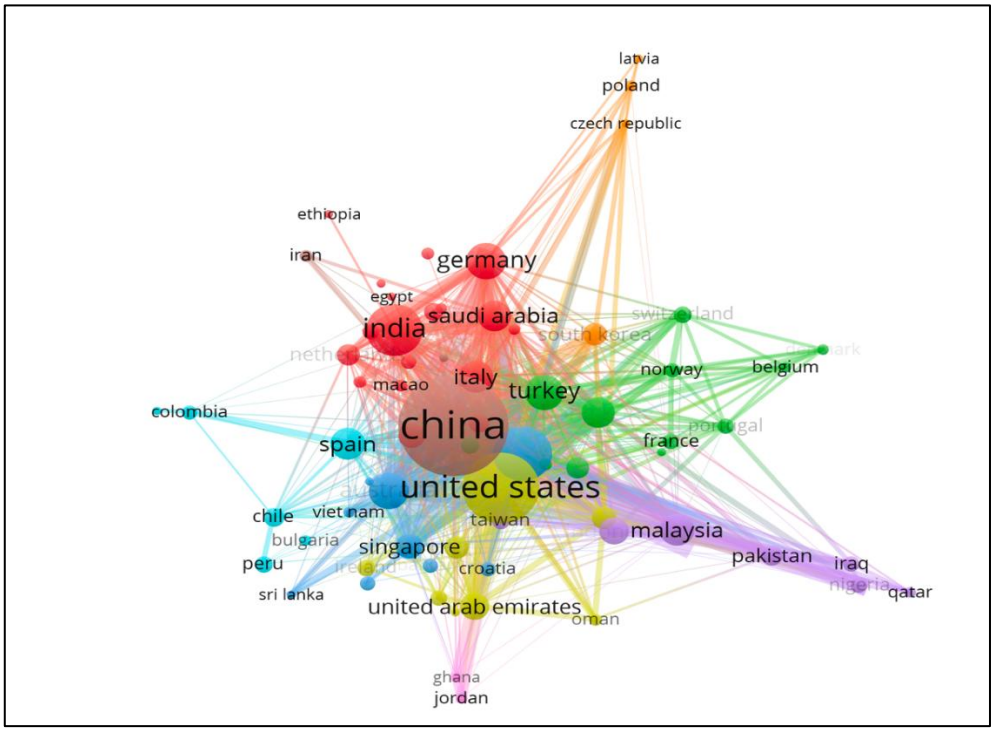

Figure 6: Bibliographic coupling by country

The bibliographic coupling analysis at the country level (Table 5, Figure 6) reveals substantial differences in research productivity, citation impact, and intellectual connectivity within the field of AI-driven assessment in education. China emerges as the most strongly coupled country, producing 212 documents, receiving 2,095 citations, and recording the highest total link strength (TLS = 6,176). Closely following is the United States with 106 documents, 3,889 citations, and TLS = 6,016. While the United States achieved the highest citation count, China's slightly stronger coupling suggests a greater degree of shared intellectual foundations with other countries in the network. This finding reflects China's growing prominence in AI and educational technology research and its increasing integration within global knowledge networks (Leydesdorff & Wagner, 2008).

The United Kingdom ranks third (63 documents; 1,021 citations; TLS = 5,519), indicating a strong role in shaping the field's intellectual structure despite producing fewer publications than China and the United States. Together, these three countries form the core of the global research network and serve as major knowledge hubs. Their central positions in the coupling map indicate extensive overlap in cited literature and strong participation in international scholarly conversations.

Beyond these dominant contributors, several countries demonstrate noteworthy coupling strengths relative to their publication outputs. Malaysia (23 documents; TLS = 2,937) and Australia (31 documents; TLS = 2,793) exhibit strong intellectual connectivity despite more modest publication volumes. This pattern suggests active engagement with internationally recognised research streams and highlights their roles as important regional contributors. Similarly, Hong

Kong (21 documents; 1,118 citations; TLS = 2,487) demonstrates exceptionally high citation impact relative to its output, indicating the influence of its publications within the broader literature. Such patterns are characteristic of strategically positioned research systems that facilitate international knowledge exchange (Wagner et al., 2017).

India (58 documents; TLS = 2,323), Germany (31 documents; TLS = 2,214), Saudi Arabia (22 documents; TLS = 2,208), and Türkiye (29 documents; TLS = 2,164) form a second tier of influential contributors. Their strong coupling strengths indicate increasing integration into the global AI assessment research landscape. The network visualisation further shows that these countries occupy important bridging positions connecting different regional clusters, reflecting the growing internationalisation of AI-related educational research (Adams, 2013).

Overall, three key dynamics emerge. First, major research hubs such as China, the United States, and the United Kingdom dominate the field in terms of productivity, citation impact, and intellectual influence. Second, countries such as Malaysia, Australia, and Hong Kong serve as important connectors, strengthening knowledge exchange across regions. Third, emerging contributors including India, Saudi Arabia, Germany, and Türkiye are becoming increasingly integrated into the global research ecosystem. Consistent with Glänzel and Schubert (2004), the findings demonstrate that intellectual influence is determined not only by publication volume or citation counts but also by the degree to which countries share and contribute to common knowledge bases. As AI-driven assessment continues to evolve, international collaboration will remain essential to addressing shared challenges in assessment validity, academic integrity, fairness, and governance.

3.4 Synthesis of Findings Relevant to the Assessment Framework

The findings from the publication trend analysis, keyword co-occurrence analysis, and bibliographic coupling analysis collectively demonstrate that AI-driven assessment research has evolved from a primarily technological focus toward broader concerns relating to assessment design, human judgement, and governance. The rapid growth in publications between 2023 and 2025 reflects increasing scholarly attention following the emergence of generative AI tools such as ChatGPT and Gemini. At the same time, the co-occurrence analysis revealed that keywords associated with assessment practices, evaluation, critical thinking, AI literacy, and generative AI have become central themes within the literature. Furthermore, the bibliographic coupling analyses identified a growing convergence on shared intellectual foundations, particularly regarding assessment redesign, academic integrity, and responsible AI use.

Taken together, these findings suggest that the future of AI-driven assessment is not determined solely by technological capability. Rather, effective implementation depends on three interrelated dimensions: (1) the design of valid and meaningful assessment tasks, (2) the allocation of responsibilities between humans and AI systems, and (3) governance mechanisms that ensure fairness, accountability, and transparency. These dimensions repeatedly emerged across

the various bibliometric analyses and therefore form the basis of the proposed Human-AI Governance and Design (HAGD) Framework.

3.4.1 Assessment design

Assessment design constitutes the pedagogical core of the framework. The co-occurrence analysis revealed that assessment-related concepts, including evaluation, assessment practices, critical thinking, academic writing, and student engagement, occupy central positions within the knowledge network. Notably, assessment serves as a linking concept connecting AI technologies to teaching, learning, and educational outcomes. This suggests that the primary challenge is no longer whether AI can support assessment, but how assessment can be redesigned to remain valid and meaningful in AI-rich learning environments.

Recent studies increasingly advocate context-sensitive, authentic, and formative assessment approaches that emphasise higher-order thinking and constructive alignment in response to generative AI capabilities (Cope et al., 2021; Mao et al., 2024; Lye & Lim, 2024; Zhai & Nehm, 2023). Consequently, task type and context, feedback alignment, and validity considerations form the framework's foundational layer. The Assessment Purpose and Validity component is conceptually aligned with Messick's (1989) unified validity framework and Kane's (2013) argument-based approach to validation, which emphasise the alignment between assessment interpretation, intended use, and evidentiary support.

3.4.2 Human–AI role allocation

The second layer concerns the allocation of responsibilities between AI systems and human assessors. The prominence of keywords such as generative artificial intelligence, automated feedback, ChatGPT-4, and Google Gemini, together with the strong coupling observed among influential authors and documents, indicates growing consensus around hybrid assessment models. Rather than replacing educators, AI is increasingly positioned as a support tool that enhances efficiency, consistency, and feedback generation.

However, highly coupled works such as Rudolph (2023), Smolansky (2023), Nikolic et al. (2023), Nikolic et al. (2024), and Floden (2025) consistently emphasise the continued importance of human judgement for interpretation, ethical reasoning, contextual evaluation, and decision-making. Empirical evidence therefore suggests that effective AI-driven assessment requires transparent allocation of roles between human and machine agents rather than complete automation (Usher, 2025).

3.4.3 Governance and oversight

The outer layer represents governance and oversight mechanisms. Across the co-occurrence network, keywords such as cheating, AI literacy, evaluation, and critical thinking reflect growing concerns regarding academic integrity, bias, accountability, and responsible AI use. Furthermore, the country-level bibliographic coupling analysis revealed increasing international engagement in AI-driven assessment research, suggesting that governance challenges extend

across institutional and national boundaries. Despite the growing importance of these issues, governance-related discussions remain fragmented and unevenly developed across the literature. Consequently, policy and accountability, fairness and bias mitigation, transparency and explainability, and ethical and regulatory alignment are positioned as overarching safeguards that guide both assessment design and human-AI interactions (Zawacki-Richter et al., 2019; Giannakos et al., 2025).

3.4.4 Human-AI governance and design (HAGD) framework

Based on the synthesis of publication trends, keyword co-occurrence patterns, and bibliographic coupling analyses, the Human-AI Governance and Design (HAGD) Framework is proposed as an integrated approach to AI-driven assessment in education. The framework was developed from three recurring themes identified across the literature: (1) the need for valid and meaningful assessment design, (2) the growing role of human-AI collaboration in assessment processes, and (3) increasing concerns regarding governance, accountability, and responsible AI use. Together, these themes suggest that effective AI-driven assessment requires balancing technological capabilities with pedagogical principles and institutional safeguards.

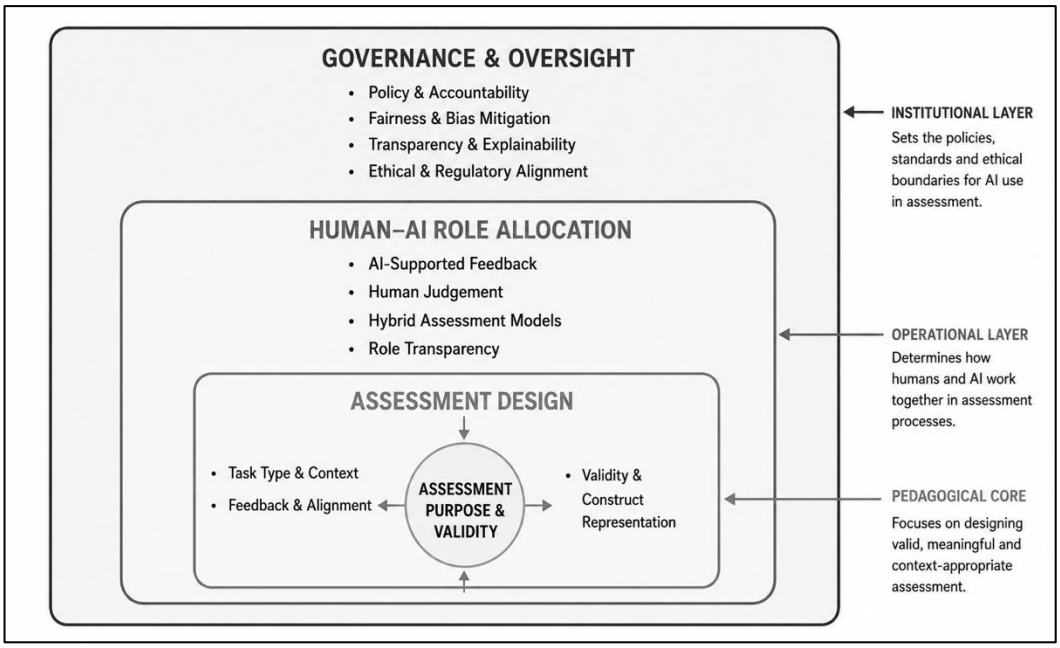

Figure 7: Human-AI governance and design (HAGD) framework for AI-driven assessment in education

The HAGD Framework (Figure 7) conceptualises AI-driven assessment as a nested socio-technical system comprising three interdependent dimensions. At the core, Assessment Purpose and Validity guide assessment design decisions relating to task type and context, feedback alignment, and construct representation. Surrounding this core, Human-AI Role Allocation determines how responsibilities are distributed between AI systems and human assessors, including the use of AI-supported feedback, hybrid assessment models, and professional judgement. The outermost dimension, Governance and Oversight,

provides institutional safeguards through policy and accountability, fairness and bias mitigation, transparency and explainability, and ethical and regulatory alignment. For example, in a ChatGPT-assisted essay assessment, Assessment Design determines learning outcomes and task authenticity; Human–AI Role Allocation assigns preliminary feedback to AI and final judgement to educators; Governance and Oversight ensure transparency and academic integrity.

The nested structure signifies dependency rather than hierarchy. Assessment design serves as the pedagogical foundation for human-AI interactions, while governance mechanisms influence and regulate both assessment design and role allocation decisions. Consequently, the framework positions AI-driven assessment not as a purely technological innovation but as a socio-technical system in which assessment validity, human expertise, and responsible governance must operate in concert to ensure educational quality, fairness, and public trust.

4. Conclusion, Limitations and Future Direction

This bibliometric study provides a comprehensive overview of AI-driven assessment research in education, based on 820 Scopus-indexed publications from 2015 to 2025. The findings show that artificial intelligence is reshaping educational assessment not only through technological innovation but also through its influence on how assessment is designed, implemented, and governed. While no publications were identified between 2015 and 2017, research activity emerged in 2018 and expanded rapidly after 2023, largely driven by generative AI technologies such as ChatGPT and Gemini.

Co-occurrence analysis revealed that the field has evolved beyond technical discussions of AI systems toward broader concerns relating to assessment practices, critical thinking, academic integrity, AI literacy, and human–AI collaboration. Bibliographic coupling analysis further demonstrated growing intellectual convergence around assessment redesign, responsible AI use, and governance mechanisms, suggesting that the field is gradually maturing into a more coherent body of knowledge.

A key contribution of this study is the development of the Human–AI Governance and Design (HAGD) Framework, comprising three interdependent dimensions: Assessment Design, Human-AI Role Allocation, and Governance and Oversight. The framework conceptualises AI-driven assessment as a socio-technical system that balances technological capabilities with pedagogical principles, human expertise, and institutional safeguards.

Nevertheless, the study is limited by its reliance on the Scopus database and the inherent constraints of bibliometric analysis, which identify research patterns but do not evaluate the effectiveness of AI-driven assessment practices. Future research should focus on empirical validation of the HAGD Framework, longitudinal and cross-cultural investigations, and the examination of governance mechanisms, fairness, and hybrid human-AI assessment models to ensure that AI-

supported assessment remains educationally meaningful, ethically responsible, and aligned with established assessment principles.

Data availability: All data generated or analysed during this study are included in this published article.

Conflict of interests: The author(s) declare no competing interests.

Ethical statements: Ethics approval is not required.

Acknowledgements: Not available.

5. Author contributions

All authors contributed to the development of this manuscript. Author 1 conceptualised the study, designed the methodology, and led the writing process. Authors 2 and 3 contributed to data collection, analysis, and interpretation. Author 4 supported the literature review and data visualisation. Author 5 was involved in reviewing and editing the manuscript. All authors read and approved the final manuscript.

6. Use of AI

The authors used ChatGPT (OpenAI) and Grammarly during the preparation of this manuscript to assist with language editing, grammar correction, sentence refinement, and improving overall readability. These tools were employed solely to enhance the manuscript's clarity and presentation, particularly because the corresponding author is not a native English speaker. The AI tools were not used to generate research data, conduct analyses, interpret findings, develop the methodology, or formulate conclusions. All research activities, including study design, data collection, bibliometric analyses, interpretation of results, framework development, and final manuscript preparation, were undertaken and verified by the authors, who assume full responsibility for the content and integrity of the work.

7. References

Adams, J. (2013). The fourth age of research. *Nature, 497*(7451), 557-560. https://doi.org/10.1038/497557a

Awidi, I. T. (2024). Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence, 6*, 100226. https://doi.org/10.1016/j.caeai.2024.100226

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389-2404. https://doi.org/10.1002/asi.21419

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access, 8*, 75264-75278. https://doi.org/10.1109/ACCESS.2020.2988510

Cope, B., Kalantzis, M., & Searsmith, D. (2021). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educational Philosophy and Theory, 53*(12), 1229-1245. https://doi.org/10.1080/00131857.2020.1728732

Fagbohun, O., Iduwe, N. P., Abdullahi, M., Ifaturoti, A., & Nwanna, O. M. (2024). Beyond traditional assessment: Exploring the impact of large language models on grading practices. *Journal of Artificial Intelligence, Machine Learning and Data Science, 2*(1), 1-8. https://doi.org/10.51219/JAIMLD/oluwole-fagbohun/19

Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal, 51*, 201–224. https://doi.org/10.1002/berj.4069

Gazni, A., & Didegah, F. (2016). The relationship between authors' bibliographic coupling and citation exchange: Analyzing disciplinary differences. *Scientometrics, 107*(2), 609-626. https://doi.org/10.1007/s11192-016-1856-y

Glänzel, W., & Czerwon, H. J. (1995). A new methodological approach to bibliographic coupling and its application to research-front and other core documents. In *Proceedings of the Fifth International Conference of the International Society for Scientometrics and Informetrics* (pp. 167-176).

Glänzel, W., & Schubert, A. (2004). Analysing scientific networks through co-authorship. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 257-276). Springer. https://doi.org/10.1007/1-4020-2755-9_12

Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leo, D., ... & Rienties, B. (2025). The promise and challenges of generative AI in education. *Behaviour & Information Technology, 44*(11), 2518-2544. https://doi.org/10.1080/0144929X.2024.2394886

Halkiopoulos, C., & Gkintoni, E. (2024). Leveraging AI in e-learning: Personalized learning and adaptive assessment through cognitive neuropsychology - A systematic analysis. *Electronics, 13*(18), 3762. https://doi.org/10.3390/electronics13183762

Hamid, H., Zulkifli, K., Naimat, F., Che Yaacob, N. L., & Ng, K. W. (2023). Exploratory study on student perception on the use of chat AI in process-driven problem-based learning. *Currents in Pharmacy Teaching and Learning, 15*(12), 1017-1025. https://doi.org/10.1016/j.cptl.2023.10.001

Harzing, A. W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics, 106*(2), 787-804. https://doi.org/10.1007/s11192-015-1798-9

Hooda, M., Rana, C., Dahiya, O., Rizwan, A., & Hossain, M. S. (2022). Artificial intelligence for assessment and feedback to enhance student success in higher education. *Mathematical Problems in Engineering, 2022*, Article 5215722. https://doi.org/10.1155/2022/5215722

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign. https://curriculumredesign.org/wp-content/uploads/AIED-Book-Excerpt-CCR.pdf

Imran, M., & Almusharraf, N. (2024). Google Gemini as a next generation AI educational tool: A review of emerging educational technology. *Smart Learning Environments, 11*(1), 22. https://doi.org/10.1186/s40561-024-00310-z

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics, 1*(4), 287-307. https://doi.org/10.1016/j.joi.2007.07.004

Jukiewicz, M. (2024). The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process. *Thinking Skills and Creativity, 52*, 101522. https://doi.org/10.1016/j.tsc.2024.101522

Khlaif, Z. N., Ayyoub, A., Hamamra, B., Bensalem, E., Mitwally, M. A. A., Hattab, M. K., & Shadid, F. (2024). University teachers' views on the adoption and integration of generative AI tools for student assessment in higher education. *Education Sciences, 14*(10), 1090. https://doi.org/10.3390/educsci14101090

Klarin, A. (2024). How to conduct a bibliometric content analysis: Guidelines and contributions of content co-occurrence or co-word literature reviews. *International Journal of Consumer Studies, 48*(2), e13031. https://doi.org/10.1111/ijcs.13031

Lelescu, A., & Kabiraj, S. (2024). Digital assessment in higher education: Sustainable trends and emerging frontiers in the AI era. In G. Grosseck, S. Sava, G. Ion, & L. Mălita (Eds.), *Digital assessment in higher education* (Lecture Notes in Educational Technology). Springer. https://doi.org/10.1007/978-981-97-6136-4_2

Leydesdorff, L., & Wagner, C. (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics, 2*(4), 317-325. https://doi.org/10.1016/j.joi.2008.07.003

Liu, R. L. (2017). A new bibliographic coupling measure with descriptive capability. *Scientometrics, 110*(2), 919-935. https://doi.org/10.1007/s11192-016-2200-2

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson.

Lye, C. Y., & Lim, L. (2024). Generative artificial intelligence in tertiary education: Assessment redesign principles and considerations. *Education Sciences, 14*(6), 569. https://doi.org/10.3390/educsci14060569

Mao, J., Chen, B., & Liu, J. C. (2024). Generative artificial intelligence in education and its implications for assessment. *TechTrends, 68*, 58-66. https://doi.org/10.1007/s11528-023-00911-4

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education and Macmillan.

Molenaar, I. (2022). The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning. *Computers and Education: Artificial Intelligence, 3*, 100070. https://doi.org/10.1016/j.caeai.2022.100070

Mohamad Hanefar, S.B., Benaouda, B., Faizuddin, A. *et al.* Mapping the Landscape of Spiritual Intelligence: A Bibliometric Analysis of Trends, Patterns and Future Directions. *Journal of Religion and Health, 64*, 3419-3447 (2025). https://doi.org/10.1007/s10943-025-02386-4

Naidu, K., & Sevnarayan, K. (2023). ChatGPT: An ever-increasing encroachment of artificial intelligence in online assessment in distance education. *Online Journal of Communication and Media Technologies, 13*(3), e202336. https://doi.org/10.30935/ojcmt/13291

Nikolic, S., Daniel, S., Haque, R., Neal, P., & Sandison, C. (2023). ChatGPT versus engineering education assessment: A multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education, 48*(4), 559-614. https://doi.org/10.1080/03043797.2023.2213169

Nikolic, S., Sandison, C., Haque, R., Hassan, G. M., & Neal, P. (2024). ChatGPT, Copilot, Gemini, SciSpace and Wolfram versus higher education assessments: An updated multi-institutional study of the academic integrity impacts of generative artificial intelligence (GenAI) on assessment, teaching and learning in engineering. *Australasian Journal of Engineering Education, 29*(2), 126-153. https://doi.org/10.1080/22054952.2024.2372154

Oc, Y., Gonsalves, C., & Quamina, L. T. (2025). Generative AI in higher education assessments: Examining risk and tech-Savviness on students' adoption. *Journal of Marketing Education*, *47*(2), 138-155. https://doi.org/10.1177/02734753241302459

Prasad, R. D., Lim, S. P., Che Yob, F. S., Wong, Y. V., Magulod, G. C., Jr., & Adom, D. (2025). Navigating the tech turn: A bibliometric analysis of decision-making trends in 21st-century education. *International Journal of Learning, Teaching and Educational Research, 24*(11), 297–313. https://doi.org/10.26803/ijlter.24.11.14

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3*, 121-154. https://doi.org/10.1016/j.iotcps.2023.04.003

Roca, M. D. L., Chan, M. M., Garcia-Cabot, A., Garcia-Lopez, E., & Amado-Salvatierra, H. (2024). The impact of a chatbot working as an assistant in a course for supporting student learning and engagement. *Computer Applications in Engineering Education, 32*(5), e22750. https://doi.org/10.1002/cae.22750

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching, 6*(1), 342-363. https://doi.org/10.37074/jalt.2023.6.1.9

Smolansky, A., Cram, A., Raduescu, C., Zeivots, S., Huber, E., & Kizilcec, R. F. (2023). Educator and student perspectives on the impact of generative AI on assessments in higher education. In *Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23)* (pp. 378-382). Association for Computing Machinery. https://doi.org/10.1145/3573051.3596191

Sweeney, S. (2023). Who wrote this? Essay mills and assessment – Considerations regarding contract cheating and AI in higher education. *International Journal of Management Education, 21*(2), 100818. https://doi.org/10.1016/j.ijme.2023.100818

Tossell, C. C., Tenhundfeld, N. L., Momen, A., Cooley, K., & De Visser, E. J. (2024). Student perceptions of ChatGPT use in a college essay assignment: Implications for learning, grading, and trust in artificial intelligence. *IEEE Transactions on Learning Technologies, 17*, 1069-1081.

Usher, M. (2025). Generative AI vs instructor vs peer assessments: A comparison of grading and feedback in higher education. *Assessment & Evaluation in Higher Education, 50*(6), 1-16. https://doi.org/10.1080/02602938.2025.2487495

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538. https://doi.org/10.1007/s11192-009-0146-3

Vashishth, T. K., Sharma, V., Sharma, K. K., Kumar, B., Panwar, R., & Chaudhary, S. (2024). AI-driven learning analytics for personalized feedback and assessment in higher education. In T. V. T. Nguyen & N. T. M. Vo (Eds.), *Using traditional design methods to enhance AI-driven decision making* (pp. 206-230). IGI Global. https://doi.org/10.4018/979-8-3693-0639-0.ch009

Villegas-Ch, W. E., Govea, J., Gutierrez, R., & Mera-Navarrete, A. (2024). Improving interaction and assessment in hybrid educational environments: An integrated approach in Microsoft Teams with the use of AI techniques. *IEEE Access, 12*, 93723-93738.

Vittorini, P., Menini, S., & Tonelli, S. (2021). An AI-based system for formative and summative assessment in data science courses. *International Journal of Artificial Intelligence in Education, 31*, 159-185. https://doi.org/10.1007/s40593-020-00230-2

Wagner, C. S., Whetsell, T. A., & Leydesdorff, L. (2017). Growth of international collaboration in science: Revisiting six specialties. *Scientometrics, 110*(3), 1633-1652. https://doi.org/10.1007/s11192-016-2230-9

Xiao, Y., & Watson, M. (2019). Guidance on conducting a systematic literature review. *Journal of Planning Education and Research, 39*(1), 93-112. https://doi.org/10.1177/0739456X17723971

Xu, H., Gan, W., Qi, Z., Wu, J., & Yu, P. S. (2024). Large language models for education: A survey. *arXiv*. https://doi.org/10.48550/arXiv.2405.13001

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? *International Journal of Educational Technology in Higher Education, 16*(39), 1-27. https://doi.org/10.1186/s41239-019-0171-0

Zhai, X., & Nehm, R. H. (2023). AI and formative assessment: The train has left the station. Journal of Research in Science Teaching, 60(6), 1390-1398. https://doi.org/10.1002/tea.21885