Copyright © 2026 The Author(s)
Didaktika: Jurnal Kependidikan Vol. 15 No. 4, November 2026

Mapping Generative AI in STEM Assessment and Feedback: A Bibliometric Study

Siti Imas Masitoh[1], I Made Astra[2], Dian Handayani[3], Firmanul Catur Wibowo[4]
[1,2,3,4]Universitas Negeri Jakarta, Indonesia

[1]siti_1310825006@mhs.unj.ac.id
DOI: https://doi.org/10.58230/27454312.5282

Abstract

This study examines how research on generative artificial intelligence (GenAI) in STEM assessment and feedback is distributed and conceptually organized. The study analyzed 181 publications dated 2023-2026 from Scopus and the Web of Science Core Collection using bibliometrix and Biblioshiny. Annual output rose from four publications in 2023 to ten in 2024, 60 in 2025, and 107 in the incomplete 2026 observation period. The corpus comprised 108 harmonized publication venues and 652 authors, with international co-authorship of 8.84%. Applied Sciences and Computers and Education: Artificial Intelligence each contributed nine publications. The author-keyword network displayed 40 nodes in four Louvain communities connecting GenAI technologies, disciplinary education, assessment, feedback, and learner-related concepts. In the All Keywords thematic map, large language models occupied the motor-theme quadrant and personalized learning the niche-theme quadrant. These findings characterize the organization of published research; they do not establish instructional effectiveness. Database coverage, incomplete current-year observation, lexical variation, and gaps in screening documentation limit the interpretation. The resulting map supports further investigation of disciplinary assessment validity, misconception-sensitive feedback, feedback uptake, delayed transfer, and human oversight.
Keywords: bibliometric analysis; feedback; formative assessment; generative artificial intelligence; STEM education

Introduction

In science, technology, engineering, and mathematics (STEM) education, an assessment should reveal more than whether a learner has obtained the expected answer. A correct numerical result may conceal a flawed model, whereas an incorrect result can expose a productive if unfinished line of reasoning. Feedback is therefore consequential when it helps learners distinguish task accuracy from the quality of the underlying process (Hattie & Timperley, 2007). Generative artificial intelligence (GenAI) has complicated this distinction: large language models can propose tasks, comment on solutions, and produce fluent explanations even when disciplinary reasoning is unreliable (Yan et al., 2024). The issue is not simply whether an AI-generated comment is plausible, but whether its judgment corresponds to what the assessment is intended to measure and can be acted upon by the learner (Carless & Boud, 2018).

Educational applications of artificial intelligence predate ChatGPT. Zawacki-Richter et al. (2019) reviewed AI-based tutoring, prediction, assessment, and personalized support, highlighting limited educator involvement in system development. Ouyang & Jiao (2021) distinguished AI-directed, AI-supported, and learner-centered arrangements. These distinctions matter for assessment: assigning a score, locating a misconception, and guiding the revision of

an explanation serve different instructional functions. They require different evidence of student understanding and different forms of human oversight.

The release of conversational GenAI intensified these debates. Lo (2023) reviewed early educational applications, whereas Tlili et al. (2023) studied ChatGPT as a case of educational chatbot use. Farrokhnia et al. (2024) and Yan et al. (2024) examined opportunities alongside institutional and ethical risks. Automated feedback, however, has a longer history (Cavalcanti et al., 2021). Its use in STEM introduces specific requirements involving equations, units, graphs, experimental procedures, and programming code. Bewersdorff et al. (2023) compared LLM judgments of errors in student experimentation protocols with human ratings. Pepin et al. (2025) in turn, considered ChatGPT in mathematics education, including its implications for assessment and task design. Together, these studies suggest that accuracy, disciplinary alignment, and opportunities to use feedback deserve separate examination.

Existing reviews address educational AI broadly (Zawacki-Richter et al., 2019), automated feedback across online learning environments (Cavalcanti et al., 2021), or a particular STEM discipline (Pepin et al., 2025). Their contributions do not, by themselves, show how the intersection of GenAI, assessment or feedback, and STEM education is represented across publication venues and research keywords. Bibliometric mapping addresses this narrower question by describing scientific output, citation visibility, and associations among indexed concepts (Donthu et al., 2021). The present analysis does not assess intervention effectiveness or infer the absence of empirical studies from a low keyword frequency. Rather, it identifies which research concerns are connected in the retrieved corpus and where targeted empirical comparisons could be informative.

We examine 181 publications dated 2023-2026 through three research questions. RQ1 asks how publication activity, citation visibility, and contributions vary across years, sources, authors, and institutions. RQ2 examines how keyword associations and thematic positions characterize the corpus's conceptual structure. RQ3 considers the geographical distribution and international collaboration patterns recorded in the metadata. Together, these questions connect performance indicators with science mapping. We interpret the resulting patterns alongside prior educational research to formulate further research directions, while keeping bibliographic observations distinct from claims about learning outcomes.

Method

This study adopted a descriptive bibliometric design that combines performance analysis with science mapping (Aria & Cuccurullo, 2017; Donthu et al., 2021). Each indexed publication formed one unit of analysis. Performance indicators describe annual output, publication venues, authors, affiliations, countries, and citation counts. Science mapping addresses keyword co-occurrence, thematic centrality and density, and country-level co-authorship. Keyword networks represent conceptual associations, whereas country co-authorship networks represent recorded collaboration. We do not report co-citation or bibliographic coupling results. Interpretation remains tied to publications and their metadata: network centrality depends on graph construction, weighting, and node selection, and it does not establish the scientific validity or causal effects of an educational intervention.

This study retrieved bibliographic records from Scopus and the Web of Science Core Collection on 4 October 2026. The retained publications span 2023-2026. Performance and thematic outputs were exported from Biblioshiny on 11 October 2026; the final author-keyword visualization was exported on 12 October 2026. Citation values refer to the retrieved metadata used for analysis, rather than a new citation search on either export date. The 2026 publication

Copyright © 2026 The Author(s)

count covers an incomplete year. As a bibliometric study, the analysis concerns publication records and does not involve participant recruitment or a classroom intervention.

The unit of analysis was the indexed publication record, not a participant, classroom intervention, or independent empirical study. The original exports contained 181 Scopus records and 86 Web of Science records (267 entries in total). Bibliometrix removed 76 duplicated entries, yielding 191 unique records. The historically retained dataset comprised 181 publications. Document-type tags identified 158 article-category records (including early-access and data-paper variants) and 23 review-category records (including one early-access review). These publication classifications cannot be interpreted as counts of independent interventions or participants.

The database searches combined three concepts: GenAI or large language models, educational assessment or feedback, and STEM teaching or learning. Scopus used TITLE-ABS-KEY and Web of Science used Topic (TS). The corpus was restricted to English-language publications dated 2023–2026, including articles and reviews and their composite indexation tags. Screening used titles, abstracts, author keywords (DE), and indexed keywords (ID). Operational relevance required a substantive GenAI element, a clear assessment/feedback concern, and an identifiable STEM education setting. These criteria defined the intended scope; keyword matches served as triage indicators rather than conclusive evidence of substantive eligibility. Studies using an LLM as an assessment benchmark were not excluded solely on the grounds that the model was not delivering feedback to learners.

Figure 1 summarizes the documented, PRISMA-informed record accounting (Page et al., 2021). First, Bibliometrix merged the two exports and removed duplicates. Secondary checks searched for repeated DOIs and normalized titles; a Jaro–Winkler similarity threshold of 0.95 was used only to identify possible title matches for inspection. The 191 unique records were then examined through automated text matching against GenAI, assessment/feedback, and STEM term lists across title, abstract, and keyword metadata. The script assigned conceptual flags and prioritized records for review, after which the researcher confirmed the historical 181-record dataset. Comparison of the archived 191-record file and the 181-record analytical file established that ten identifiable records were not retained. The original screening worksheets were initialized as UNSCREENED and do not preserve a verified reason for each non-retention decision; therefore, no article-specific grounds for exclusion or separate full-text appraisal are asserted. A later metadata audit considered five of these ten records potentially eligible under the stated scope. We preserve the actual 181-publication analytical corpus and disclose the resulting risk of selection bias rather than retrospectively rewriting the original decisions.

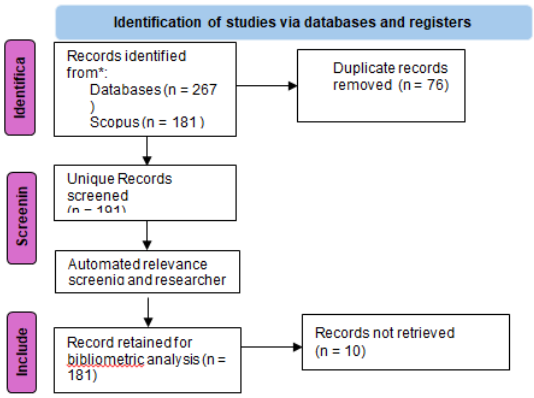


Figure 1. PRISMA-informed bibliographic record flow

We conducted performance analysis in R 4.6.1 using bibliometrix 5.5.0 and Biblioshiny, with RStudio supporting the analytical workflow. The harmonized source export combines the Applied Sciences title variants into a single venue and contains 108 sources. For keyword co-occurrence, we selected Author Keywords (DE), applied association-strength normalization, and used Louvain clustering with automatic layout. The node-selection parameter was set to 50, the minimum edge-frequency threshold to 2, and community repulsion to 0.5. We removed isolated nodes and disabled node coloring by publication year. The resulting visualization displays 40 nodes in four communities. The word cloud and single-period thematic map use All Keywords, combining author and indexed terms. The thematic map retained 250 words, a minimum cluster frequency of five per thousand documents, an alpha parameter of 0.5, Louvain clustering, community repulsion of 0.5, and three labels per theme at label size 0.3. Country collaboration used the Countries field, association-strength normalization, Louvain clustering, automatic layout, a 50-node selection, and a minimum edge-frequency threshold of one. The final country display removed isolated nodes, used repulsion of 0.5, and retained publications with more than 20 authors. Its filtered node set differs from the workbook's broader centrality table. Country frequencies, corresponding-author counts, and international co-authorship percentages represent different units and are interpreted separately. The search strings are reproduced below, and keyword-field distinctions are specified here for reproducibility.

Scopus search expression: TITLE-ABS-KEY(("generative artificial intelligence" OR "generative AI" OR GenAI OR ChatGPT OR "large language model*" OR "generative pre-trained transformer*" OR "generative pretrained transformer*" OR "GPT-3" OR "GPT-3.5" OR "GPT-4" OR "GPT-4o" OR "GPT-5") AND ("educational assessment" OR "formative assessment" OR "summative assessment" OR "authentic assessment" OR "assessment design" OR "assessment redesign" OR "assessment practice*" OR "assessment valid*" OR "automated assessment" OR "AI-generated assessment" OR "AI-assisted assessment" OR "automated grading" OR "assessment feedback" OR "formative feedback" OR "automated feedback" OR "AI-generated feedback" OR "AI-assisted feedback" OR "personalized feedback" OR "personalised feedback" OR "instructional feedback" OR "evaluative feedback" OR "peer feedback" OR "feedback literacy" OR "feedback uptake") AND ("STEM education" OR "STEM learning" OR "science education" OR "science learning" OR "science teaching" OR "mathematics education" OR "mathematics learning" OR "math education" OR "engineering education" OR "engineering learning" OR "technology education" OR "computer science education" OR "physics education" OR "chemistry education" OR "biology education"))

Web of Science search expression: TS=(("generative artificial intelligence" OR "generative AI" OR GenAI OR ChatGPT OR "large language model*" OR "generative pre-trained transformer*" OR "generative pretrained transformer*" OR "GPT-3" OR "GPT-3.5" OR "GPT-4" OR "GPT-4o" OR "GPT-5") AND ("educational assessment" OR "formative assessment" OR "summative assessment" OR "authentic assessment" OR "assessment design" OR "assessment redesign" OR "assessment practice*" OR "assessment valid*" OR "automated assessment" OR "AI-generated assessment" OR "AI-assisted assessment" OR "automated grading" OR "assessment feedback" OR "formative feedback" OR "automated feedback" OR "AI-generated feedback" OR "AI-assisted feedback" OR "personalized feedback" OR "personalised feedback" OR "instructional feedback" OR "evaluative feedback" OR "peer feedback" OR "feedback literacy" OR "feedback uptake") AND ("STEM education" OR "STEM learning" OR "science education" OR "science learning" OR "science teaching" OR "mathematics education" OR "mathematics learning" OR "math education" OR "engineering education" OR "engineering learning" OR "technology education" OR "computer science education" OR "physics education" OR "chemistry education" OR "biology education"))

Copyright © 2026 The Author(s)

The full search strategies for Scopus and Web of Science are provided above. Figure 1 presents the documented record counts, while the accompanying text explains how these figures differ from subsequent assessments of publication eligibility.

Results

Corpus Composition and Growth of Scientific Production

Table 1 summarizes the 181-publication corpus, which includes 108 harmonized venues and 652 authors. Biblioshiny reported 25 single-authored publications and an average of 3.73 co-authors per document. The bibliographic summary lists 9,409 cited references, 591 distinct author-keyword labels, and 668 indexed-keyword labels. Recorded citations total 1,070, equivalent to an average of 5.912 per document after rounding. From the endpoints of four publications in 2023 and 107 in 2026, the compound annual growth rate is 199.07% across three intervals. This calculation is sensitive to the small starting value and the incomplete final year; it should not be interpreted as a stable estimate of future growth.

Table 1. Bibliometric characteristics of the analyzed corpus

Indicator	Value
Time span	2023–2026
Documents / harmonized sources	181 / 108
Authors / single-authored documents	652 / 25
Co-authors per document	3.73
Cited references / author keyword labels	9,409 / 591
Average citations per document	5.912
International co-authorship	8.84%

Annual publication counts rose from four in 2023 to ten in 2024 and 60 in 2025, followed by 107 in the incomplete 2026 observation period (Figure 2). The last two years account for 167 of the 181 publications, or 92.3%, and the 2026 snapshot alone accounts for 59.1%. The distribution shows that recent publications dominate this corpus. It provides a measure of research output within the search boundaries, without establishing that instructional effectiveness improved alongside publication activity.

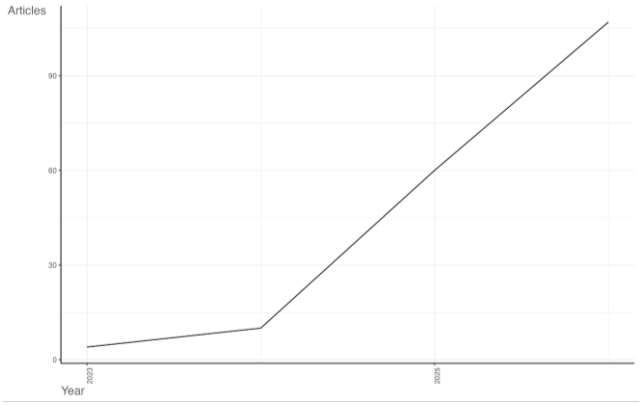


Figure 2. Annual scientific production in the 181-document corpus, 2023–2026. The 2026 observation is incomplete.

Publication Sources, Authors, and Institutional Contributions

Educational technology, engineering education, and multidisciplinary venues contribute to the source distribution. Applied Sciences and Computers and Education: Artificial Intelligence each published nine corpus records, followed by Computer Applications in Engineering Education with eight (Table 2). Four venues contributed five records each: British Journal of Educational Technology, IEEE Access, International Journal of Artificial Intelligence in Education, and International Journal of Learning, Teaching and Educational Research. Harmonization combines Applied Sciences (Switzerland) and Applied Sciences-Basel, whose five and four records refer to one nine-publication venue. Keeping these raw labels separate would divide that venue in the source ranking.

Table 2. Leading publication sources after source-title harmonization

Publication source	Documents
Applied Sciences	9
Computers and Education: Artificial Intelligence	9
Computer Applications in Engineering Education	8
British Journal of Educational Technology	5
IEEE Access	5
International Journal of Artificial Intelligence in Education	5
International Journal of Learning, Teaching and Educational Research	5

Within the corpus, the source-impact export gives British Journal of Educational Technology an h-index of five. Computers and Education: Artificial Intelligence and International Journal of Artificial Intelligence in Education each have an h-index of four; Applied Sciences has an h-index of three across its nine publications. These values derive from selected publications and their recorded citation counts, so they differ from journal-wide h-indices and from local citation measures. The author-productivity export contains several ties at two publications. Among affiliation labels, University System of Georgia has nine records, East China Normal University seven, and Monash University six. University of Georgia also appears separately in the same export. Institutional systems and constituent campuses consequently cannot be treated as independent, fully harmonized university rankings.

Global Citation Visibility

Table 3 identifies the five highest global-citation records in the retrieved bibliographic snapshot. Lee & Zhai (2024) lead with 106 citations, followed by Pham et al. (2023) with 89; the remaining values are 57, 57, and 56. The DOI, rather than the abbreviated Biblioshiny display label, identifies each publication. Counts represent the database snapshot used for the study. They are not local citations within the 181-record dataset, and they are not standardized for publication age.

Copyright © 2026 The Author(s)

Table 3. Five publications with the highest recorded global citation counts

Biblioshiny document label	Total citations	DOI
Lee & Zhai (2024), IEEE T Learn Technol	106	10.1109/TLT.2024.3401457
Pham et al. (2023), Australas J Educ Technol	89	10.14742/ajet.8825
Bewersdorff et al. (2023), Comput Educ: Artif Intell	57	10.1016/j.caeai.2023.100177
Pepin et al. (2025), Digit Exp Math Educ	57	10.1007/s40751-025-00172-1
Ilieva et al. (2025), Information	56	10.3390/info16060472

Geographical Output and International Collaboration

The country-production export reports affiliation-derived frequencies of 50 for the United States, 40 for China, 17 for the United Kingdom, 16 for Germany, and 13 for South Korea. These frequencies are not mutually exclusive national publication totals. Corresponding-author counts use a different unit: China accounts for 27 publications, comprising 26 single-country and one multiple-country publication, while the United States accounts for 26, comprising 24 single-country and two multiple-country publications. International co-authorship is 8.84% according to Biblioshiny. Each measure describes a different aspect of geographical participation. Combining them into one national ranking would obscure their different denominators and counting rules.

After isolated nodes were hidden, the country collaboration display retained a larger connected component and several smaller components (Figure 3). The larger component includes the United States, China, Canada, Singapore, South Korea, and other country labels; one smaller component connects Malaysia and Indonesia. A country can contribute publications without appearing in this filtered display. The archived centrality table uses a broader node set, and country-name variants further affect network positions. Consequently, its numerical centrality values to Figure 3, nor are they used to rank national research leadership.

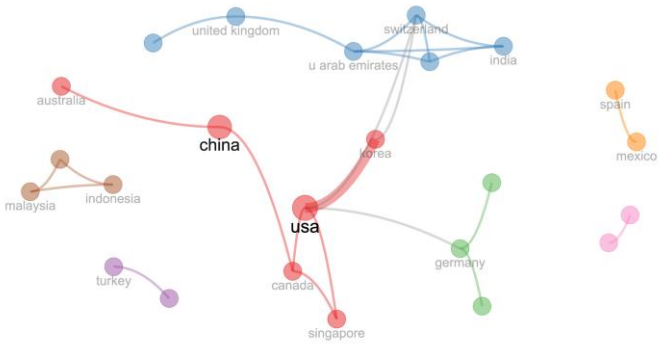


Figure 3. Country-level co-authorship network from the archived Biblioshiny analysis of 181 publications.

Keyword Co-occurrence and Conceptual Organization

The author-keyword co-occurrence graph contains 40 displayed terms assigned to four Louvain communities (Figure 4). The green community includes ChatGPT, engineering education, academic integrity, authentic assessment, and assessment validity. The blue community connects generative AI, STEM education, personalized and adaptive learning, formative feedback, and physics education. The red community includes large language models, artificial intelligence, programming education, computer science education, and automated feedback. The purple community places feedback and formative assessment alongside mathematics education and learning analytics. These are co-occurring author labels, not classifications of intervention effectiveness.

Technological expressions and educational terms intersect across the displayed communities. Several lexical variants nevertheless remain as separate nodes, including generative AI, GenAI, and generative artificial intelligence, as well as large language models and large language models (LLMs). This incomplete term harmonization affects the observable structure and prevents treating every label as a distinct research construct.

Figure 5 shows the relative prominence of terms in the All Keywords field. Engineering education, artificial intelligence, student- and teaching-related expressions, and language-model terms are prominent. Because All Keywords combines author and indexed vocabulary, its term set is broader than the DE-only network in Figure 4. Word-cloud size measures representation in the selected metadata, not educational effectiveness or theoretical importance.

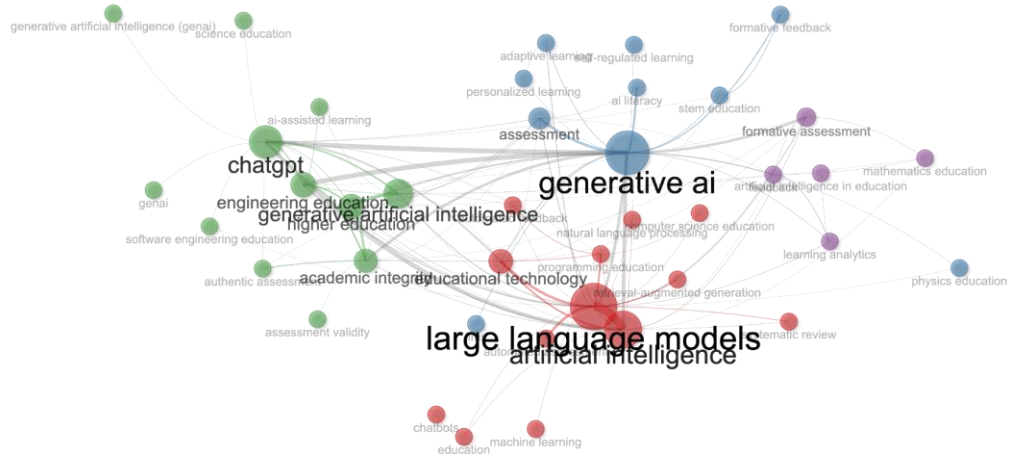


Figure 4. Author-keyword co-occurrence network for the 181-publication corpus, displaying 40 nodes in four Louvain communities.

The analysis used association-strength normalization, automatic layout, a node-selection setting of 50, a minimum edge-frequency threshold of 2, and community repulsion of 0.5. Isolated nodes were removed. Colors distinguish communities rather than publication years. Supplementary Table S4 lists the displayed labels.

Copyright © 2026 The Author(s)



Figure 5. Word cloud of the All Keywords representation in the 181-publication corpus. Font size reflects relative term prominence within this export.

Thematic Centrality and Density

The All Keywords thematic map contains five thematic groups (Figure 6). Large language models, learning systems, and language model occupy the motor-theme quadrant, where centrality and internal density are relatively high. Personalized learning, automation, and natural language processing systems lie in the niche-theme quadrant. Engineering education, artificial intelligence, and generative AI fall near the centrality boundary in the lower-density region. Students, teaching, and education computing appear near the quadrant boundaries, whereas education, systematic literature review, and science education occupy the lower-left region. That region can contain emerging or declining themes. A single-period map cannot distinguish between those trajectories without additional temporal evidence.

The thematic map in Figure 6 was generated from All Keywords, whereas Figure 4 draws only on author-assigned terms. The corresponding network communities and thematic groups are therefore not directly interchangeable. The map represents a single observation period and does not establish whether a low-centrality theme is emerging or declining.



Figure 6. All Keywords thematic map for the 181-publication corpus. Centrality indicates external thematic connectedness, and density indicates internal cohesion. Dashed lines divide the displayed thematic regions; positions describe a single observation period.

Discussion

The concentration of publications in 2025-2026 indicates a rapid increase in research attention within the selected corpus. Earlier educational studies document interest in conversational GenAI (Lo, 2023; Tlili et al., 2023), but this bibliometric analysis does not test why output increased. Database coverage, indexing delays, and the partial 2026 observation period also shape the distribution. Evaluating whether AI-generated feedback improves scientific reasoning, supports independent problem solving, or retains its benefits after assistance ends requires evidence about interventions and learning outcomes. Publication growth alone cannot answer those questions.

Publication across educational technology, engineering, and computing venues complicates attempts to generalize the meaning of assessment. Executable tests may be useful for evaluating code, yet they are insufficient for judging whether a student's physical model is conceptually defensible. A laboratory explanation can also be scientifically incomplete despite grammatically polished prose. Cavalcanti et al. (2021) describe considerable variation in automated feedback systems; Bewersdorff et al. (2023) offer a disciplinary example in which LLM error detection is compared against human assessment. Read together with the present source distribution, these studies indicate why feedback validation should specify the task, rubric, error category, and subject-specific standard rather than rely on a single generic accuracy indicator.

The author-keyword communities place GenAI technologies near assessment validity, academic integrity, individualized learning, and feedback. Co-occurrence does not demonstrate that the corresponding studies share a theoretical model. A useful interpretive distinction emerges, however, when the bibliometric pattern is considered alongside feedback scholarship. Hattie & Timperley (2007) differentiate feedback directed to tasks, processes, and self-regulation; Carless & Boud (2018) argue that feedback becomes educationally consequential through learners' capacity to interpret and use it. Applied to GenAI-supported STEM assessment, these perspectives separate three questions: Is the automated judgment scientifically warranted? Can learners understand and act on the feedback? Does revised performance transfer to subsequent tasks without assistance? The co-occurrence map identifies where these concerns appear in the literature, while the theories clarify what bibliometric links alone cannot establish.

Personalized learning occupies a relatively specialized position in the thematic map, although thematic location is not evidence of pedagogical maturity. What counts as personalization also differs: a system may adjust difficulty, generate hints, change conversational tone, or diagnose errors in a discipline-specific response. Pepin et al. (2025) discuss assessment and task design in mathematics education, where the nature of a mathematical argument matters as much as the final answer. Extending this concern across STEM would require comparisons among generic comments, diagnosis of representational or conceptual errors, and adaptive prompts whose support is gradually reduced. Feedback uptake and delayed transfer should be measured rather than assumed.

The bibliographic patterns and cited disciplinary studies suggest several research directions, although the map cannot independently verify evidence gaps. In physics, researchers could examine whether feedback preserves relationships among symbolic expressions, graphs, units, and physical meaning. Engineering and programming studies could compare transparent evaluation rules with expert judgments, including cases in which automated feedback confidently presents an incorrect assessment. Reporting should identify the rubric, model version, human verification procedure, and observed failure cases. Comparisons of feedback

Copyright © 2026 The Author(s)

timing, depth, and withdrawal could then test effects on self-regulation and later independent performance. Disagreement between teacher and AI assessments also warrants examination, since it can arise from model error, an ambiguous rubric, or legitimate differences in disciplinary judgment.

Responsible assessment extends beyond detecting plagiarism. Academic integrity is present in the keyword map, while Cotton et al. (2024) and Farrokhnia et al. (2024) detail educational tensions surrounding ChatGPT. A credible GenAI-supported assessment process would specify allowable AI assistance, protect student submissions, keep the original reasoning traceable, provide a route to challenge an automated judgment, and retain teacher authority over consequential decisions. These are proposed safeguards rather than outcomes measured by the bibliometric dataset. They also underline a theoretical distinction between a system's capacity to generate plausible judgments and an institution's responsibility to validate how those judgments affect students.

Geographical findings describe recorded participation and collaboration within the corpus. China and the United States have the largest corresponding-author counts, while international co-authorship accounts for fewer than one in ten publications. The larger component and smaller bilateral or regional components in Figure 3, including the Malaysia–Indonesia connection, show recorded co-authorship ties. They provide no basis for ranking national scholarly merit or educational benefit. Filtering isolated nodes changes the display, and separate United Arab Emirates labels can divide connections associated with one country. Stronger claims about brokerage would require reconciliation of those labels and alignment between the displayed graph and the broader centrality export.

These findings are bounded by database coverage, English-language restrictions, metadata quality, and the incomplete 2026 observation period. More importantly, the historical screening workbook does not document article-level reasons for the ten non-retained records. Retrospective review found that five of those records may satisfy the more explicitly articulated GenAI-assessment/feedback-STEM scope; their omission introduces a plausible selection-bias risk. The published metrics and figures consequently describe the retained 181 records, not all potentially eligible records. Additional limitations arise from variants in source, affiliation, country, and keyword metadata. The country map hides isolated nodes and contains separate United Arab Emirates labels, while the thematic map covers a single period. Citation visibility and network centrality are properties of this constructed corpus; neither demonstrates assessment quality, learner benefit, nor causality.

The contribution of this study is therefore methodological as well as substantive. By reporting publication output separately from conceptual and social networks, and by distinguishing author keywords from combined indexing terms, it makes the interpretive units of a rapidly expanding field more explicit (Aria & Cuccurullo, 2017; Donthu et al., 2021). Earlier reviews characterize the breadth of educational AI and automated feedback (Cavalcanti et al., 2021; Zawacki-Richter et al., 2019), whereas the present map locates their intersections with disciplinary assessment and GenAI terminology in one bounded corpus. Its central proposition is not that more GenAI publications imply better assessment. Rather, future empirical work should connect scientific correctness, students' use of feedback, and independent performance over time, with transparent human oversight (Bewersdorff et al., 2023; Carless & Boud, 2018).

Conclusion

Within the historical 181-publication corpus, research on GenAI, assessment, and feedback in STEM education rose sharply during 2023-2026. The 40-node author-keyword

network brought together technological, assessment-related, and disciplinary concepts in four Louvain communities. The All Keywords thematic map located language-model expressions in the motor-theme quadrant and personalized learning in the niche-theme quadrant; international co-authorship accounted for 8.84% of publications. These are indicators of research attention and bibliographic connectedness, not evidence that GenAI improves assessment or learning. The study contributes an explicitly bounded map and a conceptual agenda linking disciplinary validity, feedback uptake, learner independence, and human accountability. The absence of preserved exclusion reasons, retrospective identification of potentially eligible records, unharmonized metadata, and the incomplete 2026 observation limit generalization. Future studies should test the proposed links using validated rubrics, human comparison judgments, learner-level outcomes, and delayed assessment tasks.

References

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E., & Nerdel, C. (2023). Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *Computers and Education: Artificial Intelligence*, *5*, Article 100177. https://doi.org/10.1016/j.caeai.2023.100177

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, *2*, Article 100027. https://doi.org/10.1016/j.caeai.2021.100027

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, *61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Lee, G.-G., & Zhai, X. (2024). Using ChatGPT for science learning: A study on pre-service teachers' lesson planning. *IEEE Transactions on Learning Technologies*, *17*, 1643–1660. https://doi.org/10.1109/TLT.2024.3401457

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), Article 410. https://doi.org/10.3390/educsci13040410

Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, *2*, Article 100020. https://doi.org/10.1016/j.caeai.2021.100020

Copyright © 2026 The Author(s)

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLOS Medicine*, *18*(3), Article e1003583. https://doi.org/10.1371/journal.pmed.1003583

Pepin, B., Buchholtz, N., & Salinas-Hernández, U. (2025). A scoping survey of ChatGPT in mathematics education. *Digital Experiences in Mathematics Education*, *11*, 9–41. https://doi.org/10.1007/s40751-025-00172-1

Pham, T., Nguyen, T. B., Ha, S., & Nguyen Ngoc, N. T. (2023). Digital transformation in engineering education: Exploring the potential of AI-assisted learning. *Australasian Journal of Educational Technology*, *39*(5), 1–19. https://doi.org/10.14742/ajet.8825

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, *10*, Article 15. https://doi.org/10.1186/s40561-023-00237-x

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112. https://doi.org/10.1111/bjet.13370

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*, Article 39. https://doi.org/10.1186/s41239-019-0171-0

--- This page has been intentionally left blank ---