Article:

Ogunleye, B., Olayinka, O., Zakariyyah, K.I. et al. (2026) An analysis of AI-enabled authentic assessment strategy using text mining techniques. Review of Education, 14 (2). e70206. ISSN: 2049-6613

https://doi.org/10.1002/rev3.70206

Received: 26 November 2025 | Accepted: 20 July 2026

DOI: 10.1002/rev3.70206

REVIEW AND SYNTHESIS

Review of Education BERA

An analysis of AI-enabled authentic assessment strategy using text mining techniques

Bayode Ogunleye[1] | **Olakunle Olayinka**[2] | **Kudirat Ibilola Zakariyyah**[1] | **Oluwaseun Ajao**[3] | **Hemlata Sharma**[4]

[1]School of Architecture, Technology & Engineering, University of Brighton, Brighton, UK

[2]School of Computer Science, The University of Sheffield, Sheffield, UK

[3]Department of Computing & Mathematics, Manchester Metropolitan University, Manchester, UK

[4]School of Information, Journalism and Communication, The University of Sheffield, Sheffield, UK

Correspondence
Bayode Ogunleye, School of Architecture, Technology and Engineering, University of Brighton, Room C523, Cockcroft Building, Lewes Road, Brighton BN2 4GJ, UK.
Email: b.ogunleye@brighton.ac.uk

Abstract

The development and use of artificial intelligence (AI) tools have raised several concerns about the effectiveness of assessment practice. Authentic assessment has emerged as one of the key strategies to cope with the rapidly evolving technology. Nevertheless, prior studies argued that authentic assessment as a standalone instrument is insufficient to control the impact of AI. Thus, several studies called for the integration of AI in assessment design. However, it is unclear how academics can deploy the use of AI in authentic assessment. This paper aims to present AI-enabled authentic assessment strategies through evidence synthesis. Following the PRISMA guidelines, we collected 103 documents from multiple databases and utilised text mining techniques, including co-occurrence analysis and topic modelling, to synthesise the documents. The results showed that academic integrity, critical thinking, framework, integration of AI, and teaching and learning methods are the dominant themes across the analysis. Most importantly, our findings identified four pathways in which AI can be deployed. Therefore, we propose the DAD (Deployment options, AI role, and Deliverables) conceptual framework to organise knowledge and guide the future design of AI-enabled authentic assessment.

KEYWORDS

artificial intelligence, authentic assessment, co-occurrence analysis, generative AI, higher education, topic modelling

Context and implications

Rationale for this study: The emergence and increasing usage of AI technologies have challenged assessment practices in higher education. While a growing number of studies have proposed various solutions, including frameworks, they address a certain discipline or isolated element, such as authenticity or reflective practice.

Why the new findings matter: Our results provide a more holistic and theoretically grounded assessment strategy addressing the alignment between AI integration, pedagogical design, and cognitive engagement to evidence learning in an assessment of learning context.

Implications for practitioners, policymakers, and researchers: For authenticity, our result emphasises the need for academics, education practitioners, and policymakers to ensure assessment of learning policies is not against the use of AI. As AI continues to evolve, our result suggests that education practitioners need to ensure that assessment design allows the use of AI as both an agent and a tool to mirror 'real-world' professional context without undermining cognitive engagement.

INTRODUCTION

Technological advancement has influenced teaching and learning pedagogical practice, more specifically, assessment, which is a key component of learning. Assessment for learning and assessment as learning both provide opportunities for formative assessment, while assessment of learning, also termed summative assessment, predominantly aims to evaluate and summarise students' performance at the conclusion of an instructional exercise, determining the degree and extent of learning achieved. Assessment, teaching, and learning are intertwined. Each forms an integral part of the pedagogical process and occurs in a cycle. This cycle is updated and upgraded through the feedback mechanisms. Within the cycle, assessment supports learning by providing students the opportunity to demonstrate skills and knowledge acquired and their application towards academic and professional competencies. Assessment helps identify students' learning progress and areas for improvement by systematically measuring their understanding and skills. It thus serves as a cornerstone for evaluating students' comprehension, identifying learning gaps, and providing actionable feedback that can significantly impact academic attainment and future success. Assessment thus remains the most significant prompt for learning (Ashford-Rowe et al., 2014; Boud, 2012; Villarroel et al., 2018). Unfortunately, recent technological developments like artificial intelligence (AI) pose challenges to assessment practices. Prior studies have shown that the AI systems, specifically, the generative artificial intelligence (GenAI), can pass assessments in data science (Ogunleye et al., 2024a), law (Choi et al., 2021), pharmacy (Kunitsu, 2023), and programming (Malinka et al., 2023) subjects, thereby increasing the potential for unethical use. Other notable problems identified include academic integrity (Tran et al., 2025), over-reliance and misuse (Lee et al., 2024), privacy and bias (Nguyen, 2025). To this end, literature suggests that these constraints could mar creativity, innovativeness, problem-solving, critical thinking, and logical analysis skills, which are core to employability (Essien et al., 2024; Ogunleye et al., 2024a). To mitigate this threat, several institutions or educational bodies came up with guidelines and new policies on the usage, including no AI usage, responsible (minimal) usage and integration of the AI system.

Despite AI's sophistication, its responses tend to obscure students' actual understanding, aside from other severely highlighted ills and negative consequences (Zhao, 2024).

This creates a significant transformation in the educational sector partly because of the critical challenges regarding validity, reliability, academic integrity and ethics, hence the need to re-image assessment by (re-)emphasising the need for competencies (not mere skill mastery), and application of knowledge that suits a concern, a project, an environment, a group, or a circumstance (not just theoretical knowledge). Traditional assessment does not always or adequately foster critical thinking or creativity (Fatima et al., 2024). Prior studies argued that real-world problems are not presented as a multiple choice question or solved based on memorisation of facts. Indeed, problem-solving is more of how facts are used with self-motivation, metacognitive, creative, analytical, and critical thinking abilities (Miserandino, 2025). This thus forms the core of authentic assessment. As technology continues to advance, the construct of authenticity in assessment needs reimagining to remain relevant in higher education (Ogunleye et al., 2024a; Vallis, 2025). This is supported by the regulatory body of Australia's higher education that recommends assessment that aligns with 'responsible and ethical use of AI in ways that are authentic to both the task and the discipline' (Tertiary Education Quality and Standards Agency, 2023).

Authentic assessment has emerged as one of the key strategies to cope with the rapidly evolving technology. Gulikers et al. (2004) defined authentic assessment as the assessment that requires students to use the same competencies, or combinations of knowledge, skills, and attitudes that they need to apply in the criterion situation in professional life. In a similar vein, Swaffield (2011) defined authentic assessment as the assessment of learning that involves the use of 'real-world tasks' requiring students to apply their knowledge and skills in meaningful contexts to solve 'real-world' problems. Villarroel et al. (2018) identified the core principles of authenticity assessment as realism, cognitive challenge, and evaluative judgement. Hamel and Lee (2024) successfully used the authentic criteria, namely, realism, cognitive challenge, and evaluative judgement, to evaluate their authentic assessment strategy. Realism implies that the assessment task is set in a 'real world' context. This allows students to solve problems similar to real-life challenges or professional tasks. The cognitive challenge involves the use of higher order cognitive skills, including critical thinking, learning, and reasoning. This helps students to develop a deeper understanding and ability to synthesise new ideas with prior knowledge and apply theoretical knowledge in practice. This is shown to improve learning and long-term retention. While the evaluative judgement involves the development of self-assessment skills through experience, processes, and feedback. This helps students to develop the ability to critique their own work, and that of others and thus promotes creativity and innovation. Authentic assessment promotes active learning, improved achievement, and greater information retention; most importantly, it provides learners with real-world hands-on experience (Murphy et al., 2017). By depicting how students arrived at and showing why they did, authentic assessment emphasises experiential learning and the importance of self-motivation, critical reflection, and analytical ability (Ashford-Rowe et al., 2014). Assessment tasks must closely mirror real-world applications, arouse integration of interdisciplinary/multidisciplinary knowledge, and complex problem-solving (in replacement for mere recall of isolated facts).

Higher education institutions have responded to the AI disruption towards assessment practice in several ways, including the creation of new policy, revised assessment format, and technological safeguards (Corbin et al., 2025). An example is the use of authentic assessment. However, authentic assessment alone is insufficient to control the impact of AI on academic integrity (Kofinas et al., 2025). Several studies have called for the integration of AI in assessment design. Xia et al. (2024) argued that there is a need to investigate how to design effective GenAI-informed assessments. This supports the evidence synthesis performed by Moorhouse et al. (2023) that suggests GenAI tools should be incorporated into assessment design. At this stage, it is unclear how to strategically deploy AI in authentic assessment. There are limited studies that showed how AI can be integrated. More importantly, there is no study found that synthesises literature on how AI can be incorporated into authentic assessment. To this end, this paper aims to analyse AI-enabled authentic assessment strategies through evidence synthesis. By doing so, we provide answers to the following research questions (RQ).

- RQ1: How are academics embedding AI into authentic assessments?
- RQ2: What AI-enabled authentic assessment frameworks exist to guide academics in the design of assessments?

To answer the research questions and fulfil the aim of this study, this paper retrieved documents from multiple databases and used text mining techniques to synthesise the documents. The rest of the paper presents the methodology (in Section 'Methodology'), results (in Section 'Results'), discussion of findings (in Section 'Discussion'), and conclusions (in Section 'Conclusion').

METHODOLOGY

This paper performed a systematic literature review (SLR) of AI-enabled authentic assessment strategies using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework (Page et al., 2021). By doing so, this study retrieved data from multiple databases using relevant keywords and applied text mining techniques for synthesis. In subsequent subsections, we present details on the data collection strategy, text mining techniques, implementation set-up, and evaluation metrics.

Data collection strategy

This study used relevance and coverage as key points in the selection of the databases to ensure global research outputs were considered. This study collected data from seven databases, namely, Scopus, Web of Science, ACM, ScienceDirect, IEEE, Taylor and Francis, and Emerald, using relevant keywords including 'artificial intelligence', 'authentic assessment' and their variants. It is worth stating that this study did not consider archive databases and grey literature. The search strings were joined using Boolean operators such as 'or', 'and' to search through the title, abstract, and keywords of publications between 2020 and June 2025. Furthermore, we made use of the asterisk (*) wild card to ensure that the keyword variations are captured and the double quotation (" ") to ensure that the exact keyword phrase is captured. However, it is worth noting that ScienceDirect does not support the use of wildcards and limits the number of Boolean connectors to a maximum of eight per field. Using the search query presented in Table 1, across the databases, our initial search produced 234 documents, which were then filtered to reflect research documents published between the year 2020 and June 2025. This yielded 223 documents. Thereafter, the documents were limited to publications in the English language and document types, articles, conference papers, book chapters, and review papers. This resulted in 204 documents. The dataset was pre-processed in the Python environment, and during this

TABLE 1 Search query.

('authentic assess*') AND (chatgpt OR 'generative AI' OR 'generative artificial intelligence' OR genai OR gai OR AI OR 'artificial intelligence') AND ('higher education' OR education OR academia or 'universit*' OR postgrad* OR undergrad* OR 'tertiary education')

stage, 52 duplicate records were removed. Moreover, we employed the inclusion and exclusion criteria as detailed in Table 2 to narrow down the documents to suit the scope of the study. This was achieved by reviewing the publication abstract in the first iteration and the full publication text in the second iteration (in cases where there is no clarity in the abstract only). During this stage, we removed 28 documents because they were not retrievable. Further to this, we removed 21 documents because they were irrelevant to the scope of the study. The documents were classed as irrelevant because the research papers focused on AI capabilities and performance on coursework/assignment without AI integration (16) and student perception of AI on coursework (5). In total, 103 documents emerged for the analysis. We have duly followed the PRISMA guidelines, and this is presented visually in Figure 1.

Data quality assessment

To ensure reliability, this paper employed two authors to manually label the documents as 'relevant' and 'not relevant', and in cases of disagreement, a third author was utilised. This study employed the use of Cohen's kappa inter-rater reliability to ensure the study's reliability. Cohen's kappa inter-rater reliability value provides an indication of the degree of consistency among raters. The Cohen's kappa coefficient is a statistical measure that depicts a value useful to evidence the extent of consistency among raters, that is, the degree to which raters' measures are the same (Cohen, 1960; McHugh, 2012). Cohen's kappa value computed yielded a kappa value of 0.83, which is a strong agreement. To ensure consistency, authors engaged in regular discussions (for review of the documents) throughout the screening process.

Text mining techniques

This study utilised term co-occurrence analysis and topic modelling (TM) to mine key information from the research publications. The co-occurrence analysis involves the identification of prevalent and key themes in research documents. A co-occurrence network is defined as a representation of the relationship between words or phrases that frequently appear together in a given domain. This study used VOSviewer, developed by Van Eck and Waltman (2017), to perform co-occurrence analysis. In VOSviewer, clustering is applied to determine the relatedness of publications. Clustering is the grouping of objects based

TABLE 2 Inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
Publications within the year 2020 and June 2025	Publications before the year 2020
Publications in the English language	Publications not in the English language
Peer-reviewed publications	Not peer-reviewed/grey literature
Publications with document types, namely, articles, conference papers, book chapters, and review papers	Editorials, meeting abstracts, workshop papers, posters, book reviews, and dissertations
Publications that are related to authentic assessment and AI	Articles related to assessment and AI, but not authentic assessment
Publications that presented or discussed the use of AI in authentic assessment	Articles related to assessment and AI, but the use of AI in authentic assessment is not presented or discussed

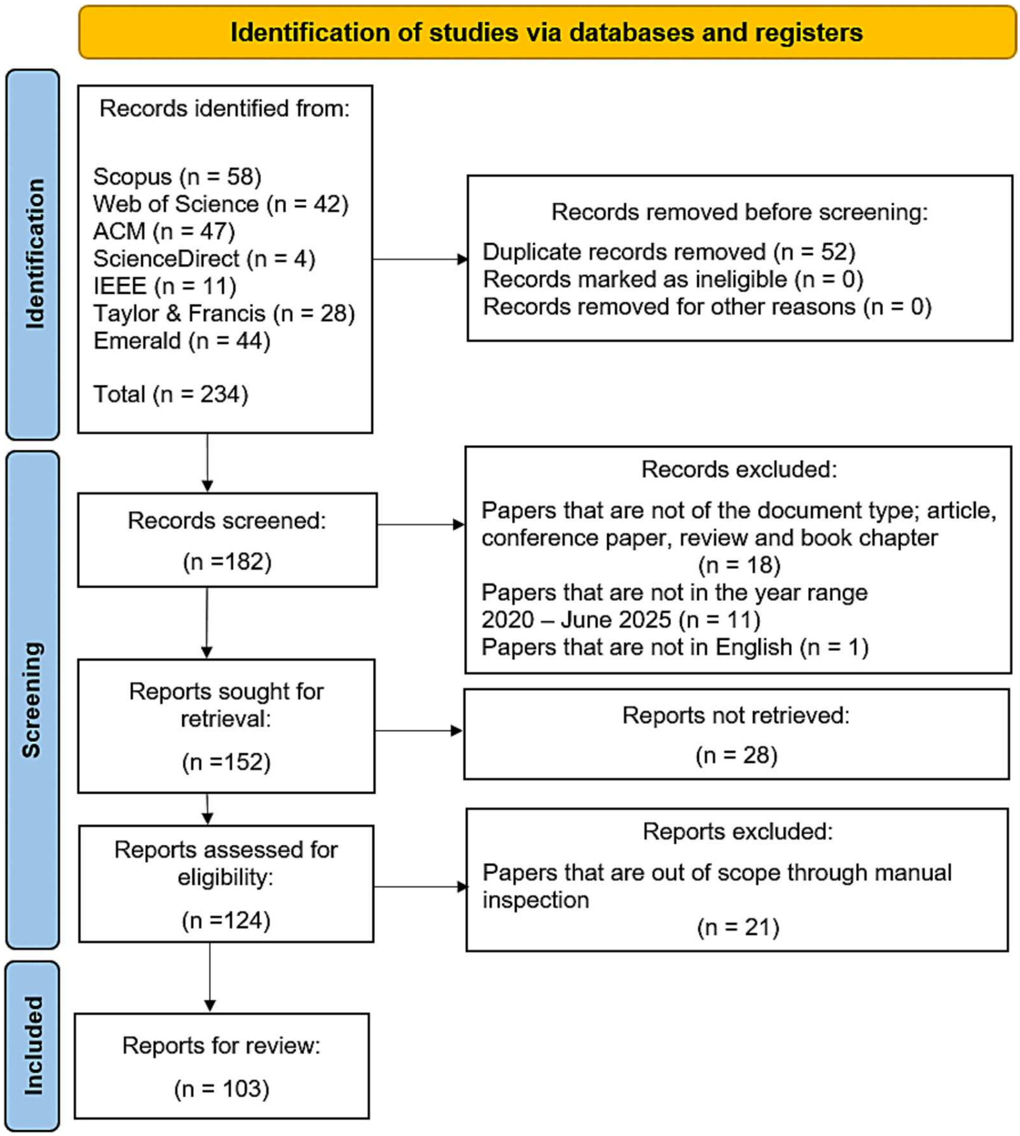

FIGURE 1 Flowchart of data collection using the PRISMA framework.

on their similar attributes. The cluster results are non-overlapping, which implies that each publication can only belong to a cluster, and there is no publication without a cluster assigned. Further to this, this study complements the co-occurrence analysis with TM. Topic modelling involves the identification of abstract themes and hidden semantic patterns in a large set of documents. Prior studies have combined these techniques for an in-depth text (semantic) analysis (Lim & Hwang, 2025; Ogunleye et al., 2025). This paper used latent Dirichlet allocation (LDA) to perform TM. This is because LDA has shown good performance across several domains (Aydin & Ogunleye, 2026; Ogunleye et al., 2023). LDA proposed by Blei et al. (2003) is a generative probabilistic model for discovering the relationships among texts in a large text corpus. LDA assumes that each document consists of a random mixture of latent topics, where each topic is characterised by a distribution over words. A basic notation is that given there are N documents, let D denote the document index, where $D = 1, 2, \ldots, N$. Then, let Z denote the topic assignment for each word in a document, where $Z_{(d,n)}$

represents the topic assignment for the nth word in document d. The model parameters α represent the parameter for the Dirichlet distribution over document-topic distributions, and β is the parameter for the Dirichlet distribution over topic-word distributions. The 'hidden' topics are unknown and unspecified before executing the LDA model. The goal of LDA is then to infer the hidden variables (Z, θ, ϕ) given the observed words (W) and the model parameters (α, β). The joint probability of the observed words W, latent topic assignments Z, and Dirichlet-distributed parameters θ and ϕ, conditioned on hyperparameters (α, β), is defined as the fully factorised distribution over topics and documents, where topic-word and document-topic distributions are drawn from Dirichlet priors, and each word is generated from a topic-specific multinomial distribution. The model works by analysing the topics while simultaneously estimating the extent to which each document in the corpus is linked to the topic.

Experimental set-up

The co-occurrence analysis was performed in VOSviewer (version 1.6.20) and visualised as an overlay and network maps. To implement the co-occurrence analysis, the dataset was imported into the VOSviewer using the 'create a map based on text data' option. The ignore structured abstract labels, ignore copyright statement, and the full counting checkboxes were ticked. A summary of the configuration is presented in Table 3. The LDA implementation was performed in the Python environment. We utilised Python packages including Gensim (Řehůřek & Sojka, 2010) and NLTK—natural language toolkit (Bird, 2006) to implement LDA and pre-processing, respectively. Beforehand, the dataset was pre-processed for punctuation expansion, stop word removal (NLTK), removal of irrelevant/noisy words, and

TABLE 3 Implementation of text mining techniques.

Implementation	Set-up
Co-occurrence analysis (VOSviewer)	Full counting, minimum number of term occurrences = 10, number of terms selected = 50, Normalisation method = association strength, minimum cluster size = 1, iterations = 10, random seed = 0, random start = 10, weight = occurrences, clustering = modularity-based clustering
LDA (Gensim)	chunksize = 1740, alpha = 'auto', eta = 'auto', random_state = 42, iterations = 1000, num_topics = 7, passes = 20, eval_every = None
Corpus	gensim.models.Phrases(noun, min_count = 20, threshold = 10), dictionary.filter_extremes(no_below = 15, no_above = 0.6)
Part of speech	NLTK-pos_tag-Noun: 'NN', 'NNPS', 'NNS', 'NNP'
Lemmatisation	NLTK package-nltk.stem.WordNetLemmatizer()
Additional words removed	'ai', 'data', 'model', 'need', 'based', 'study', 'including', 'technology', 'research', 'ha', 'method', 'wa', 'approach', 'generative', 'artificial', 'intelligence', 'generative artificial intelligence', 'artificial intelligence', 'use', 'using', 'used', 'also', '2023', 'model', 'however', 'nan', 'common', 'licence', 'copyright', 'author', 'paper', 'attribution', 'international', 'permitted', 'cc', 'taylor', 'francis', 'group', 'academic', 'higher', 'tool', 'chatgpt', 'ability', 'university', 'authenticity', 'aim', 'genai', 'focus', 'assessment', 'authentic', 'education', 'design', 'work'

lemmatisation (NLTK). The punctuation expansion helps to transform words like 'don't' into 'do not'. The stop word removal process helps to remove words that do not convey semantic meaning in the sentence. For example, 'and', 'is'. These words can dominate topic distribution in the LDA model without contributing meaningful differentiation. The removal of irrelevant and additional (domain-specific) words was performed to reduce noise or skewness within the corpus. This, in turn, helps to improve topic distinctiveness. Lemmatisation is the transformation of word forms that belong to the same inflectional morphological paradigm to their corresponding canonical form. An example is 'running', 'ran' gets converted to 'run'. Furthermore, we used the NLTK package to implement the part-of-speech tagging and thus extracted nouns (and noun phrases) to develop our corpus. This process helps to extract important terms from the corpus. Table 3 presents the details of the co-occurrence analysis and LDA implementation.

Evaluation metrics

This study used multiple metrics, including coherence scores (UMass and CV), perplexity, and human interpretability (qualitative measure) for evaluation of the LDA model. The metrics are discussed as follows:

- Topic coherence measures the semantic similarity between words within a topic and is widely regarded as a reliable indicator of model quality (Ogunleye et al., 2023). The coherence scores can range from values of 0–1, with higher values indicating more semantically coherent and interpretable topics. Coherence score (CV) measures the semantic similarity between high-scoring (most important) words in a topic. It uses normalised pointwise mutual information (NPMI), word embedding, and sliding windows to capture word co-occurrence patterns. The higher the better, as a high value is an indicator of more interpretable and meaningful topics. Coherence score (UMass) measures the co-occurrence of words within a topic based on document frequency statistics. It is based on the document co-occurrence counts and computed using a log-likelihood estimation from a corpus. A higher value is a better indicator.
- Perplexity is the measure that aids in identifying how well the LDA model predicts unseen data. It is based on a probability distribution to find out how 'surprised' the model is when encountering new documents. A lower value indicates better generalisation.
- Human interpretation is the manual examination of the quality and meaningfulness of the generated topics. It is also known as the qualitative validation of topic meaningfulness (Grimmer & Stewart, 2013).

RESULTS

This study presents our results in two ways. The first part of the result discusses the findings from the term co-occurrence mapping. While the latter part discusses the results of the LDA model.

In overlay visualisation, the colour of a term is determined by the year of the term, and the year of a term is calculated as the average publication year of all documents in which the term occurs. The results, as shown in Figure 2, evidence that active learning, academic integrity, assessment design, capability, ChatGPT, critical thinking, framework and generative AI tool are the dominant themes in the research documents. More specifically, the results show that ChatGPT, generative AI tool, assessment design, question, capacity, student learning, and learner are the dominant themes in the year 2023. This suggests that after the release of

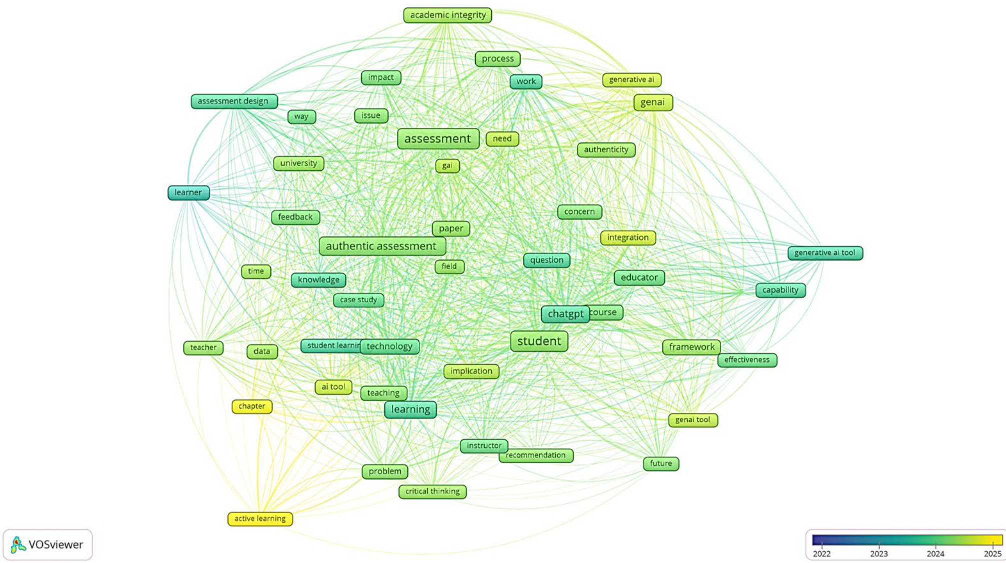

FIGURE 2 Overlay visualisation of term co-occurrence map.

ChatGPT by OpenAI on the 30th of November 2022, researchers began to investigate the use of ChatGPT for learning. Moreover, the result further suggests that researchers examined the capacity or performance of the generative AI tools on assessment questions. From the year 2024, implication, active learning, concern, framework, genai, integration, issue, academic integrity, and critical thinking are the emerging themes. This suggests that studies explored the integration of GenAI, frameworks and active learning. Furthermore, the results also suggest that research papers explored the issue, concern, and impact of GenAI. To this end, it is unsurprising to see themes including 'academic integrity' and 'critical thinking'. To have a better understanding of these themes, we produced the co-occurrence network map of the terms in Figure 3.

The term co-occurrence network map in Figure 3 shows that the research documents are grouped into five clusters. The individual clusters can be seen in Appendix A1 (Figures 3a–3e). The results show that cluster 1 (as shown in Figure 3a) consists of terms, namely, course, educator, student, capacity, effectiveness, performance, generative AI tool, integration, framework, recommendation, and future. These terms suggest that cluster 1 consists of publications on the effectiveness and integration of GenAI tools. Similarly, cluster 2 (as shown in Figure 3b) consists of terms, namely, critical thinking, authentic assessment, assessment design, active learning, problem, learning, data, knowledge, feedback, and learner. The results suggest documents in this cluster discuss the issue of 'critical thinking' and thus, discuss learning strategies including active learning, authentic assessment, feedback, and assessment design. Moreover, clusters 3, 4, and 5 consist of varied documents. For example, cluster 3 (as shown in Figure 3c) can be profiled as teaching, learning and technology because it consists of terms, namely, paper, need, gai, work, concern, way, student learning, technology, teaching, and field. Similar to this, cluster 4 (as shown in Figure 3d) can be profiled as impact, case study, and assessment question. This is because cluster 4 consists of terms, namely, teacher, AI tool, instructor, implication, ChatGPT, question, case study, time, university, and impact. While cluster 5 (as shown in Figure 3e) can be an issue for academic integrity because cluster 5 consists of terms, namely, authenticity, assessment, genai, process, issue, and academic integrity. In summary, the co-occurrence network analysis of the research (abstract) documents revealed that implication, active learning, concern,

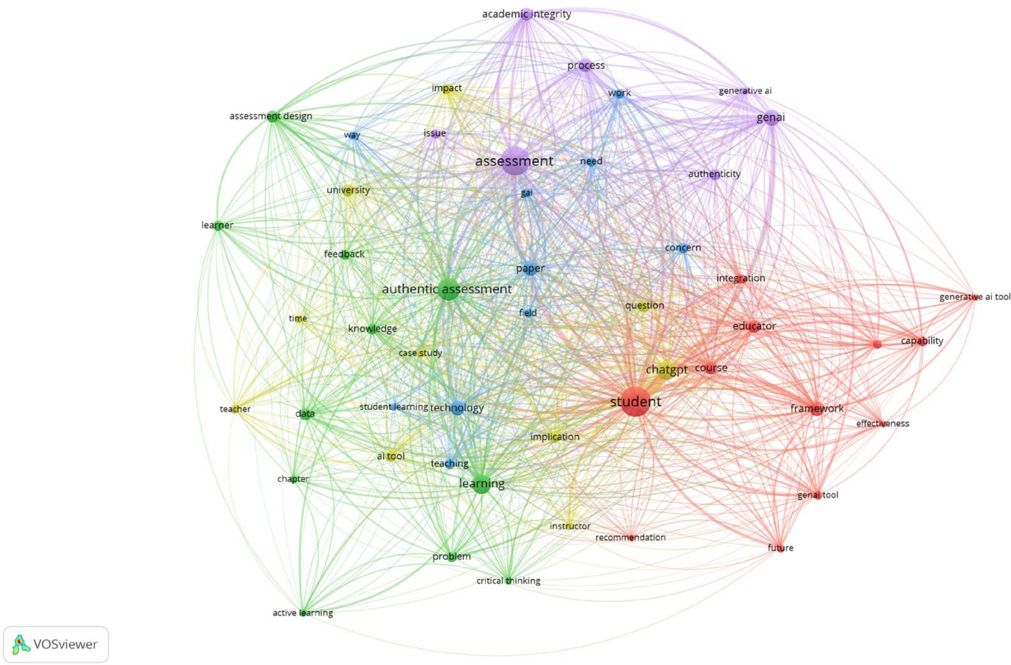

FIGURE 3 Co-occurrence network map of terms.

framework, genai, integration, issue, academic integrity, and critical thinking are the emerging themes, while effectiveness and integration of AI (cluster 1), critical thinking and learning strategies (cluster 2), teaching, learning and technology (cluster 3), impact, case study, and assessment question (cluster 4), issue and academic integrity (cluster 5) are the five cluster profiles produced. To enable more detailed analysis and complement the co-occurrence analysis, this study implements and presents topic model results in the subsequent section.

Topic modelling

A crucial step in the LDA configuration is to determine the optimal value for alpha (α), Beta (β), and the number of topics (k). This study set the hyperparameters α and β as 'auto' to enable the model to learn optimal asymmetric priors directly from the data during training. This process helps to control the sparsity and distribution of topics and words. To determine the optimal value of k, we produced the coherence scores (CV) in Figure 4 (see Appendix A2), coherence scores (UMass) in Figure 5 (see Appendix A2) scores and perplexity scores in Figure 6 (see Appendix A2) for the number of topics (k) ranging from 3 to 30. The coherence scores (CV) result shows that the model achieved optimality at $k=7$ with a value of 0.32. This implies that $k=7$ yields the highest semantic consistency among topic words relative to all other evaluated topics. The model produced the most coherent and interpretable topics at $k=7$. Coherence score (CV) has been shown in prior studies (Lim & Lauw, 2024; Ogunleye et al., 2024b; Röder et al., 2015) as the metric that correlates with human interpretability, among others. Beyond $k=7$, the coherence score (CV) declines or fluctuates without exceeding this peak, indicating that increasing the number of topics introduces topic fragmentation and reduces semantic clarity. This behaviour suggests that additional topics do not capture new meaningful structure but instead divide coherent themes into less interpretable subcomponents. The coherence score (UMass) shows marginal improvements at certain higher values of k (for example, $k=9$ and $k=17$). However, these differences are relatively small and do not outweigh the topic interpretability and stability provided by the coherence (CV) metric. Similarly, higher values of k produce redundant and noisy topics. The model exhibits overly broad topics, which limits the usefulness. The perplexity score decreases as the number of k increases. This trend is unsurprising, as perplexity tends to favour more complex models with a larger number of topics. We examined the topics produced when k is set at 5, 6, 7, 8, 9, and 10 to determine the most interpretable terms. This study opted for a 7-topic configuration based on optimal coherence score (CV) and human interpretability. This paper summarises the evaluation scores of the 7-topic LDA model in terms of coherence score (CV and UMass), and perplexity in Table 4.

LDA topics

The results from the 7-topic LDA model are presented in Table 5. The results indicate 7 distinct topics with 15 terms per topic. The terms are listed in order of their importance within each topic. In Figure 7 (see Appendix A2), we produced the terms with their weight. This, coupled with human interpretation, has been used to profile the topics for better understanding. In the subsequent section, this paper discusses each of the topics. This was achieved by thoroughly reviewing the research documents with high contributions within each topic.

Topic 1—Evaluation

Topic 1 terms like evaluation, review, enhance, support, and opportunity suggest enhancing student learning outcomes through reflection. Similarly, terms like learning, skill, knowledge, practice, thinking, and experience suggest student learning and reflection. While terms like challenge, literature, and course suggest education context. Based on these themes, Topic 1 can be profiled as ‘evaluation’. Topic 1 consists of publications that discussed how authentic assessment can be deployed by incorporating AI output evaluation with elements of reflection. For example, Miserandino (2025) presented AI-enabled authentic assessments to promote realism and develop cognitive, creative, critical thinking, subject knowledge, and problem-solving skills. Miserandino's assessment strategy incorporated students' evaluative judgement of AI outputs. The strategies were classed into five categories, namely, ‘content knowledge and application’, ‘scientific inquiry and critical thinking’, ‘values in psychological science’, ‘communication, literacy and technology’, and ‘personal and professional development’. The idea of content knowledge and application is to create questions around a particular subject or concept. Thereafter, students are required to prompt AI for a solution and then evaluate the AI output. The scientific inquiry and critical thinking involve using prompting strategies to invoke a contrarian thinking of a situation or to critique an assumption of a concept. The values in

TABLE 4 Evaluation of the LDA model.

Evaluation metrics	Score
Coherence score (CV)	0.3242
Coherence score (UMass)	−1.6446
Perplexity	−3.5232

TABLE 5 LDA topics identified.

Topic	Terms	Topic label
1	Evaluation, challenge, experience, enhance, skill, opportunity, process, review, practice, thinking, support, literature, knowledge, learning, course	Evaluation
2	Framework, skill, implication, analysis, challenge, implementation, development, environment, thinking, course, enhance, practice, result, knowledge, opportunity	Framework and skill development
3	Learning, thinking, concern, opportunity, experience, integrity, environment, enhance, application, implication, task, support, skill, question, literature	Teaching and learning methods
4	Practice, challenge, task, process, implementation, support, application, review, environment, literature, opportunity, analysis, issue, development, result	Review of practice
5	Application, response, challenge, skill, knowledge, practice, concern, opportunity, question, course, task, result, learning, thinking, process	Application of knowledge
6	Course, question, experience, development, process, educator, response, practice, result, framework, support, task, challenge, knowledge, environment	Course-specific assessment design
7	Integrity, educator, practice, challenge, issue, analysis, opportunity, concern, development, process, framework, review, skill, implementation, task	Academic integrity

psychological science entail using multiple AI systems to generate images based on certain characters and thus provide an evaluative and reflective summary of the outputs. The communication, literacy and technology strategy involves the investigation of the potential bias of AI output. While the personal and professional development entails evaluating AI outputs towards subject-specific job applications/descriptions. Moreover, Upsher et al. (2024) proposed three AI-enabled authentic assessment strategies, namely, peer learning, process folios, and the PAIR model. The peer learning strategy involves the peer evaluation of AI output in an authentic assessment. The peer evaluation process helps refine students' critical evaluation abilities and heightens their metacognitive awareness. The portfolios are focused on laying emphasis on the learning journey over the outcome. While the PAIR model (originally developed by Acar, 2023) involves four key stages, namely, '*problem formulation*', '*AI tool selection*', '*interaction with AI*', and '*reflection/evaluation*'. In the problem formulation stage, learners are required to define the problem they aim to address and subsequently, identify and choose an AI tool for their needs by evaluating the features of multiple AI systems. In the interaction stage, students experiment with the AI tool and finally assess and report their experiences with the AI system. Similarly, Chen et al. (2025) evaluated the use of an AI-enabled authentic assessment strategy implemented as a challenge-based learning through role playing as network engineers. In the assessment, students were tasked to research on the feasibility and implications of a full transition from IPv4 to IPv6. The assessment strategy involved five key stages, namely, problem formulation and understanding of requirement, inquiry, and retrieval of information from AI (with a focus on prompt engineering), critical analysis and validation of findings, synthesis of findings and production of a solution, and presentation of the solution. Students are required to deliver an interim project progress log, which should detail the list of prompts and responses, critical analyses of AI outputs, and validation of information through credible sources. In addition, submit a 3-min video presentation followed by an oral examination. Chen's assessment strategy received positive feedback from students. Lastly, Khojasteh et al. (2025) developed an academic writing self-assessment toolkit to evaluate feedback provided by an AI system to refine students' critical evaluation abilities and heighten their metacognitive awareness.

Topic 2—Framework and skill development

Topic 2 terms like framework, implementation, development, analysis, and implication suggest framework or model implementation. Terms like skill, thinking, knowledge, practice, and course suggest skill development. While terms such as environment, challenge, result, opportunity, and enhance suggest application and improvement. Based on these themes, Topic 2 can be profiled as ‘Framework and skill development’. Topic 2 consists of publications that presented a framework and skill development model for the deployment of AI-enabled authentic assessment. For example, Donnelly and Kennelly (2024) developed the Inquiry and Academic Writing (IAW) framework for incorporating AI tools into academic writing assessment tasks. The IAW framework is in four stages, namely, *discover and plan*, *develop and revise*, *reflect and connect*, and *share learning*. The discovery and plan stage involves brainstorming (including AI suggestions) and the development of the research idea. The develop and revise stage involves working towards a draft responding to the assessment task. The reflect and connect stage involves reflecting on the learning and application experience, while the share learning stage provides the opportunity for students to conceptualise their learning for peer-presentation (in any agreed or acceptable format). Moreover, Sinnappan et al. (2025) presented the MetaC (metacognitive coursework assessment) framework to enable students to effectively leverage AI systems to improve their academic performance and cognitive skills. The focus is to enhance metacognitive skills in the context of using AI tools for written assessments. MetaC consists of components to develop key attributes, namely, planning, monitoring, and evaluation skills. The planning stage involves the objective definition in alignment with the assessment task requirement and identifying the role of AI (including brainstorming and problem-solving). The monitoring involves the synthesis of work and feedback from AI tools, peers, or instructors for improvement. While the evaluation involves the reflective analysis and learning optimisation. The framework is developed for students to develop the skills to critically assess AI outputs, improve problem-solving abilities, and be able to integrate AI tools as part of a cohesive academic strategy. Similarly, Dwight (2025) used Villarroel's model (Villarroel et al., 2018) to design an AI-enabled authentic assessment framework. The framework consists of four key steps, namely, workplace context, design assessment, judgement, and feedback.

Minagar et al. (2024) presented ALAN (assessment as learning authentic tasks for networking), a framework to automate the generation of authentic assessment tasks for students. In their study, they discussed how ALAN was implemented to promote peer learning and academic integrity. Further to this, Thanh et al. (2023) employed Bloom's taxonomy as a guiding principle to create authentic assessments in the economics education context. Thereafter, they developed a framework to examine the effectiveness of AI systems in solving authentic assessments. Moreover, Pyakurel and Souppez (2025) used both employability and global consciousness components to propose a conceptual framework for designing authentic assessment. In this era where academic integrity is of importance due to AI usage, Pyakurel & Souppez stated that the concept of the employability component is that the authentic assessment should be well aligned with the development of students' knowledge, skills, and behaviours (KSBs). While the focus of the global consciousness component is that authentic assessment should incorporate societal elements to help students be socially and globally conscious. They assert that this strategy not only mitigates the misuse of AI tools but also prepares students well for their career endeavours. Hsiao et al. (2023) developed a framework to identify design dimensions for redesigning authentic assessment. Their framework consists of six components, namely, course context and learning goals, function and focus, grading criteria, modes, authenticity, and administration. The course context and learning goals involve the analysis of how the use of digital tools aligns with the learning outcomes and impacts student learning. The choice of assessment type is dependent on the depth of cognitive skills aimed to be developed. In the function and focus stage, tutors need to determine if AI systems can help students in their assessment tasks or feedback processes. In this AI era, tutors must consider incorporating AI usage in multiple processes of the assessment deliverables for both formative and summative assessments, rather than the final assessment product/deliverable. The grading criteria discuss that grading should place greater emphasis on assessing the criteria related to the cognitive skills process of the assessment deliverable. The mode stage provides flexibility; however, tutors need to determine if a small group activity for an in-class session or an off-campus activity will help towards the learning outcomes. The authenticity indicates that tutors need to ensure that assessment follows the core principles of authentic assessment (realism, cognitive challenge, and evaluative judgement). The administration indicates that tutors need to undergo adequate training in understanding the characteristics and vulnerabilities of AI systems to promote responsible usage.

Topic 3—Teaching and learning methods

Topic 3 terms such as learning, thinking, experience, question, concern, and skill suggest reflection and metacognition. Terms such as application, task, and environment suggest hands-on experience learning. While terms such as support, enhance, opportunity, implication, and literature suggest scaffolding and research-informed practice. Based on these themes, Topic 3 can be profiled as ‘Teaching and learning methods’. Topic 3 consists of publications that discussed the teaching and learning methods employed to engage students, provide real-life problem-solving experience, and develop high-order thinking skills leading to authentic assessment. For example, Salinas-Navarro et al. (2024a) examined the integration of AI tools with authentic assessment and experiential learning to understand how the pedagogy strategy can enhance teaching and learning experience. For this purpose, they prompted AI tools for suggestions on the use of AI tools for experiential learning activities for the purpose of creating authentic assessments. They synthesised the AI outputs, and their findings suggest that AI tools are useful for generating real-life problem coursework, assessment rubrics, scenario-based role-play assessments, automated peer review and feedback, and adaptive scenario-based assessments. Therefore, they concluded that AI tools should be deployed as agents to support experiential learning for authentic assessment.

Topic 4—Review of practice

Topic 4 terms such as practice, literature, support, environment, opportunity, and development suggest scaffolding and research-informed practice. Similarly, terms such as review, analysis, and suggest evaluation. Terms like implementation, process, application, task, challenge, issue, and result suggest real-world problem-based learning. Based on these themes, Topic 4 can be profiled as 'Review of practice'. Topic 4 consists of publications that discussed practices for implementing or mitigating AI use for authentic assessment. For illustration, Ramakrishnan et al. (2025) evaluated SQL programming assessment with and without the integration of an AI tool. They examined the impact of the AI-enabled authentic assessment, and their results showed that higher achieving students showed improved performance with the AI-enabled authentic assessment. However, it was less effective for students struggling with fundamental concepts. This is unsurprising as we envisage that students with surface learning might struggle more with AI-enabled assessment. Moreover, Khan et al. (2023) discussed the successful implementation and positive student feedback of their game-based and simulation learning for an AI redundancy authentic assessment strategy. O'Riordan et al. (2025) discussed the implementation of an interactive oral assessment strategy as an opportunity to mitigate the effects of AI, more importantly, as a reliable, valid, and authentic assessment approach. Similarly, Yung et al. (2024) showed that complementing interactive oral (synchronous) with reflective assessment is a valid and reliable authentic assessment method that stimulates high levels of engagement in this AI era. The interactive oral process involves students attending a synchronous interactive oral activity, which is different from a question-and-answer presentation. These are natural conversation prompts tailored around an authentic scenario. Thus, authors including Hughes et al. (2025) recommend the deployment of synchronous (live) online presentations as an authentic assessment strategy.

Topic 5—Application of knowledge

Topic 5 terms such as application, process, task, challenge, result, response, course, suggest real-world problem-based learning, while terms such as practice, concern, opportunity, learning, thinking, question, skill, and knowledge suggest reflection and metacognition. Based on these themes, Topic 5 can be profiled as 'Application of knowledge'. Topic 5 contains documents that lay emphasis on evidencing and assessing students' application of knowledge in AI-enabled authentic assessments. From the students' perspective, Alshwiah (2024) stated that students acknowledge that AI enhanced their application of knowledge in their AI-enabled assessment task processes. Fatima et al. (2024) provided a commentary report on AI-enabled authentic assessment strategies in the medical education context. They discussed various opportunities to apply theoretical knowledge to a practical simulated patient scenario. Fatima et al. focused more on the use of virtual reality and augmented reality simulations to create patient scenarios (as a case study) for immersive practice. Thereafter, students are required to analyse and evaluate the situation to make informed decisions or treatment/diagnosis. Furthermore, Reimer (2024) discussed the use of AI as a collaborator in authentic assessment designs. Reimer's exploratory study identified key factors to consider when incorporating AI into authentic assessment designs, namely, assessment priorities, rubrics, context, layer, communication, equity, critical thinking, and reflection. The author reinforces the need to use assessment rubrics as a tool (by incorporating the criteria) to evaluate the quality of AI-generated content critically and the application of theoretical knowledge to practical scenarios. Further to this, Bannister et al. (2025) presented the development and application of the PANDORA GenAI Susceptibility Rubrics to assess the susceptibility of assessments to the undeclared use of AI tools.

Topic 6—Course-specific assessment design

Topic 6 terms such as course, educator, practice, experience, environment, development, process, framework, support suggest pedagogy design, while terms such as knowledge, challenge, question, task, response, result suggest assessment. Based on these themes, Topic 6 can be profiled as ‘Course-specific assessment design’. Topic 6 publications provide course-specific AI-enabled authentic assessment. For example, Elson et al. (2025) embedded ChatGPT into their coursework assessment. The AI-enabled assessment strategy involves the use of ChatGPT to solve conceptually challenging physics problems and also write Python code. Students are required to critically assess ChatGPT's approach at each step and guide it towards the correct solution. The results of the approach suggest that students attained comparable grades to the previous year, where AI was not embedded in the assessment. More importantly, findings showed that markers should pay attention to how students interpret and steer the AI outputs. Sakzad et al. (2024) proposed process-driven diverging assessment as an AI-enabled authentic assessment strategy. Their study argued that randomising questions (from a large pool of questions) for each student has been successful in assessing learning in an online environment. However, in this AI era, they propose a process-based assessment that uses the same set of questions for all students but randomises the input to implement the authentic assessment tasks in which AI is adopted as an assistant. Kizhakkethil and Perryman (2024) required students to compare and evaluate the search results of AI against the library database. Thereafter, they analysed students' perspective of general use of AI and incorporating AI in their authentic assessment. Their findings showed that students were impressed with the output of AI but cautious. More specifically, students were concerned about reliability, lack of transparency, and over-reliance upon the technology. They stated that there is a need for continuous critical evaluation of AI outputs and awareness of ever-changing limitations of the systems, including the inability to cite AI-generated content.

Topic 7—Academic integrity

Topic 7 terms such as integrity, concern, issue, and challenge suggest integrity and ethical issues. Similarly, terms such as educator, practice, review, analysis, and opportunity suggest review of practice. While terms such as framework, development, implementation, task, process, and skill suggest framework development. Based on these themes and the high weight of the term integrity, Topic 7 can be profiled as ‘Academic integrity’. Topic 7 consists of documents that examined the design of AI-enabled authentic assessment with emphasis on academic integrity. In this AI era, academic integrity has been identified as one of the key research areas (Qi & Zhu, 2024). Lehane and Wright (2024) performed a SLR to examine the impact of authentic assessment design on academic integrity. Their findings suggest that the use of authentic assessment improved students' engagement with the module, their employability skills, and fostered academic integrity. Furthermore, Martin et al. (2025) examined academic integrity while implementing AI-enabled authentic assessment using two student groups. Their assessment task required students to create a blog that summarises and critically evaluates (including critical reflection) an empirical research article on mental health. In the first group, students were required to use AI tools to assist with drafting the blog; thereafter, they were required to evaluate and reflect on the AI output. This first student group was named the compliant group. The second group, named the unrestricted group, were allowed the use of AI throughout the assessment process, including producing the entire assessment solution. The findings of Martin et al. suggest that the unrestricted group achieved limited success in using AI to provide the entire/adequate critical reflection. Their study showed their assessment strategy promotes engagement and critical skills development; more importantly, it provides an opportunity for maintaining assessment integrity. Davey et al. (2025) argued that AI-assisted assessment delivered as an interactive oral examination ensures academic integrity. However, it requires adequate scaffolding and implementation strategies. Rasul et al. (2024) proposed an academic integrity framework that consists of eight strategies for maintaining academic integrity in higher education. Rasul et al.'s framework focused on three stakeholders, namely, students, educators, and institutions. Their framework highlighted engaging in collaborative learning, enhancing critical thinking skills, and improving AI literacy as key strategies for empowering students for integrity. Similarly, the framework identified innovative pedagogical/assessment approaches (project-based assessment, authentic assessments, and adaptive learning) and participating in development programmes as key strategies for equipping educators for integrity. While technological infrastructure, a clear GenAI policy, and creating a culture of academic integrity were key elements for institutions.

DISCUSSION

This section addresses the two research questions as follows:

> RQ1. How are academics embedding AI into authentic assessments?

The findings from the evidence synthesis revealed four pathways of embedding AI into assessments, namely, output evaluation, brainstorming and process, experiential learning, and scenario/simulation analysis. The concept of AI output evaluation involves utilising the AI system to perform the assessment task. Thereafter, the output from the AI can then be evaluated. The evaluative judgement of the AI output can be delivered in various ways. For example, by analysing the potential bias of AI output or the use of prompt strategies to invoke a contrarian thinking of a situation or critique an assumption of a concept. The brainstorming and process pathway involves utilising the AI system in different stages of the assessment deliverable (rather than using the AI output). This implies that emphasis is laid on the learning journey over the outcome. In this way, the AI system is used for planning (development of ideas), providing feedback on draft work, and is deployed for implementation. Furthermore, the experiential learning pathway involves using AI systems to create various learning activities, and the end product can be received through journaling. The final approach we discovered in literature was the use of AI to create scenarios/simulations. This offers academics opportunities to create near-real-world scenarios that would typically be difficult to achieve without AI and test students on them. This approach also opens up opportunities for personalised learning, where each student could be assessed on unique scenarios/case studies. It also provides opportunities for the creation of various formative assessments that the students could engage with to develop their skills and learn the subject matter. Table 6 below presents our findings and maps them to relevant literature.

> RQ2. What AI-enabled authentic assessment frameworks exist to guide academics in the design of assessments?

As discussed in Sections 'Topic 1—Evaluation, Topic 2—Framework and skill development, Topic 3—Teaching and learning methods, Topic 4—Review of practice, Topic 5—Application of knowledge, Topic 6—Course-specific assessment design', the results of this study identified five key frameworks that academics could use to design AI-enabled authentic assessments, namely, ALAN (Assessment as learning authentic tasks for networking),

TABLE 6 Classification of AI-enabled authentic assessment strategies.

Strategy	Description	Literature
AI output evaluation	This involves the process where the AI tool is required to perform the same assessment task as the student. Thereafter, the AI output and the human (student) output are critically evaluated. This can be delivered in the form of a reflective report, comparative analysis, or peer reflection in a group work assessment. This process is shown to enhance the development of innovation	Antoniou et al. (2025), Farrell (2024), Kizhakkethil and Perryman (2024), Martin et al. (2025), Miserandino (2025), Upsher et al. (2024)
Brainstorming and process	This involves the active use of AI tools for assessment task processes Students are required to provide or maintain a prompt diary detailing their engagement with the AI tool. Examples of AI roles include new idea generation, enquiry/feedback on a process, paragraph development, writing partner, draft reviewer (peer review), reflection assistance through feedback, and collaborator (students as partners). The assessment can be delivered as a critical report, coursework report, reflective report, self-assessment report, portfolio, or journal	Abbasnejad et al. (2025), Elson et al. (2025), Foung et al. (2024), Hatley and Penny (2024), Ping et al. (2025), Sakzad et al. (2024), Upsher et al. (2024)
Experiential learning	This is the process where an AI tool is integrated with experiential learning activities for authentic assessment The end product can be delivered as a blog, portfolio through journaling, reflective essay, or report. Examples of AI roles include peer reviewer, reflection assistance through feedback, AI evaluation, and collaborator	Chen et al. (2025), Martin et al. (2025), Salinas-Navarro et al. (2024a, 2024b)
Scenario/simulation analysis	This process involves the use of AI to create various practical scenarios or simulated scenarios (for example, representations of characters, virtual reality, and augmented reality patient scenarios). Thereafter, students are to analyse and evaluate the situation to make informed decisions (or treatment/diagnosis)	Fatima et al. (2024), Miserandino (2025)

IAW (Inquiry and Academic Writing), MetaC (Metacognitive Prompting), improved Villarroel's model, and Bloom's taxonomy. In Table 7, we present a summary of the frameworks, and analyses of the frameworks are discussed in the next subsection.

Analysis of existing AI-enabled authentic assessment frameworks

The IAW framework developed by Donnelly and Kennelly (2024) was developed as part of a CPD course to help academics (in business and education disciplines) with the design of assessments that support students' writing by creating a stage-by-stage process for learning. With this framework, academics take students through an incremental process of learning underpinned by active learning tailored specifically for their module. The framework is primarily focused on academic writing and attempts to shift the focus of academics from simply assessing the final output to emphasising the process of learning at different stages, thus helping to avoid the inappropriate use of AI in writing. Although innovative in its own right, this approach to assessment might not be appropriate to all subject areas

TABLE 7 Summary of AI-enabled Authentic Assessment Framework.

Framework	Description	Literature
ALAN (Assessment as learning authentic tasks for networking)	A framework to automate the generation of an authentic assessment task	Minagar et al. (2024)
IAW (Inquiry and academic writing)	A framework developed to design AI-enabled authentic assessment specific to academic writing	Donnelly and Kennelly (2024)
MetaC (Metacognitive coursework assessment)	For the development of authentic assessment by integrating planning, monitoring, and evaluation skills, with a focus on AI generative tools	Sinnappan et al. (2025)
Villarroel model (updated version)	Adapted Villarroel's model for the development of AI-enabled authentic assessment	Dwight (2025)
Bloom's taxonomy	For the development of authentic assessment	Chen et al. (2025), Reimer (2024)

and often requires the academic to make changes to the delivery of existing modules. This might be difficult because making changes to existing module plans and how they are taught and assessed might not always be possible, especially in institutions where these changes need to be submitted well ahead of the module being taught (in some cases, year 1). The MetaC framework by Sinnappan et al. (2025) focused on enhancing metacognitive skills of planning, monitoring, and evaluation for effective use of AI tools in written assessment. Although validated by 21 academic experts, the framework was specifically developed to help students improve critical thinking, academic performance, and structured feedback while incorporating AI in university-level assessment. There is a general suggestion of mapping the planning, monitoring, and evaluation skills from a disciplinary perspective. However, it is not clear from the original research how this framework could be adopted for effective AI-enabled assessment development. Minagar et al. (2024) presented ALAN, a framework to automate the generation of authentic assessment tasks for computer network students. In their study, they discussed how ALAN was implemented to promote peer learning and academic integrity. Although Minagar et al. indicated that AI can be used in an assistive mode, the framework did not detail an explicit integration. Further to this, Reimer (2024) discussed the use of Bloom's taxonomy and AI as a collaborator in authentic assessment designs for social work education. Reimer identified key factors to consider when incorporating AI into authentic assessment designs, namely, assessment priorities, rubrics, context, layer, communication, equity, critical thinking, and reflection. Furthermore, Dwight (2025) improved on Villarroel's model to design an AI-enabled authentic assessment framework for cybersecurity education. The framework consists of four key steps, namely, workplace context, design assessment, judgement, and feedback. This framework focused on an iterative approach, such that once the assessment has been designed, students work on a task (creation of attack trees) based on a manual process and also using AI, then compare the results from these for a period of 4 weeks, with informal feedback provided weekly during tutorial sessions. While the generalisation of results from the study is limited in that the framework has only been used for one assignment, it does offer an interesting approach to AI-enabled authentic assessments that could potentially be used beyond simply the cybersecurity domain.

Although all the frameworks reviewed have their merits and uses, they all focus on a specific aspect of AI-enabled authentic assessment, with none providing a truly holistic view of what is possible when exploring the use of AI in assessment design. An academic seeking to embed AI into their assessment should be able to make informed decisions about how to embed AI into their assessment without having to read through various case studies and research literature. A framework that should provide academics with the entire picture of what is possible and has been successfully deployed, with the opportunity to choose from the toolkit/framework based on specific relevant aspects. AI is a rapidly developing technology, and its use and adoption in academia for assessment is growing and will change over time. Many of the existing frameworks do not fully address all that is required for an assessment to be fully AI-enabled and, in most cases, speak only to one case study, thus confining it to only a type of way that AI-enabled authentic assessments can be done. This paper presents in Table 8 an analysis of the existing framework using six dimensions, namely, scope, focus, AI-pedagogical role, AI integration options, cognitive engagement, and implementation guidance. The dimensions were adapted from prior studies (Kickbusch et al., 2025; McCauley et al., 2025; Ouyang & Jiao, 2021; Powell & Forsyth, 2024; Su et al., 2026). The scope dimension refers to the discipline the framework has been developed or applied. The focus dimension refers to the primary direction of the assessment framework. The AI-pedagogical role dimension refers to the function the AI assumes within the learning and assessment process ranging from passive to active and interactive pedagogical agent to align with learning theories and outcomes (Alqarni, 2025; McCauley et al., 2025; Salinas-Navarro et al., 2024a, 2024b). The AI integration options dimension refers to the extent and ways in which the AI is embedded within the assessment framework. This is important to demonstrate the intentional and structured integration of AI system to align assessment with real-world practices (Kickbusch et al., 2025). The cognitive engagement dimension refers to the depth of learners' intellectual involvement, encompassing higher order thinking and metacognitive processes. This is important as the assessment framework must elicit critical analysis, evaluation, and reflective interaction to overcome surface-level task completion (Kickbusch et al., 2025; Peters & Angelov, 2025). The implementation guidance and transparency dimension examines if transparency and guidance were provided on adopting the framework in practice. Lastly, this paper adopts the dimension scoring scale used in the study of McCauley et al. (2025) namely, *not addressed*, *partially addressed*, and *fully addressed* to score the existing framework in Table 8.

As given in Table 8, there is no existing framework that comprehensively addresses all dimensions. In particular, the AI integration options and scope remain underdeveloped, which motivated the development of our DAD (Deployment options, AI role, and Deliverables) conceptual framework. Our framework advances existing literature by moving beyond fragmented, discipline-specific, and tool-focused approaches towards a holistic, flexible, and pedagogically grounded model. By explicitly defining AI roles, integrating multiple learning approaches, and aligning deliverables with process-oriented assessment design, it provides a scalable strategy for the development of AI-enabled authentic assessment. To this end, this study presents in Figure 8 the DAD conceptual framework for the development of AI-enabled authentic assessment.

The proposed DAD conceptual framework highlights three key areas, namely, deployment options, AI role, and expected deliverables from students. At the deployment options stage, we identified diverse learning modalities (as given in Table 6) through which academics are embedding AI into their assessment. By providing these as an initial starting point in our framework, it serves as a lens through which academics seeking to embed AI into their assessment can start to explore what is possible. Furthermore, at the AI Role, tutors are able to determine the role that AI should play in the assessment. Our results identified solution development, AI-as-a-learning tutor (reviewer/feedback provider), in-class support (supporting learning), creator, and role playing as the main roles AI can be utilised in this stage. It is important to highlight that although some AI roles flow naturally with certain deployment options, where possible, tutors are able to choose any deployment option and choose the role that they wish AI to play (this implies one-to-many options). Finally, at the deliverable stage, tutors have to decide on the expected output that students will eventually submit to be

TABLE 8 Analysis of existing frameworks.

Dimension	ALAN (Minagar et al., 2024)	IAW (Donnelly & Kennelly, 2024)	MetaC (Sinnappan et al., 2025)	Updated Villarroel (Dwight, 2025)	Use of Bloom's taxonomy (Reimer, 2024)
Scope	Discipline-specific	Discipline-specific	General	Discipline-specific	Discipline-specific
Focus	Authentic task	Staff development	AI prompting	Authentic task	Authentic task
AI-pedagogical role	Not addressed	Partially addressed	Partially addressed	Fully addressed	Not addressed
AI integration options	Not addressed	Partially addressed	Partially addressed	Partially addressed	Partially addressed
Cognitive engagement	Not addressed	Fully addressed	Fully addressed	Partially addressed	Fully addressed
Guidance for implementation	Fully addressed	Fully addressed	Partially addressed	Partially addressed	Partially addressed

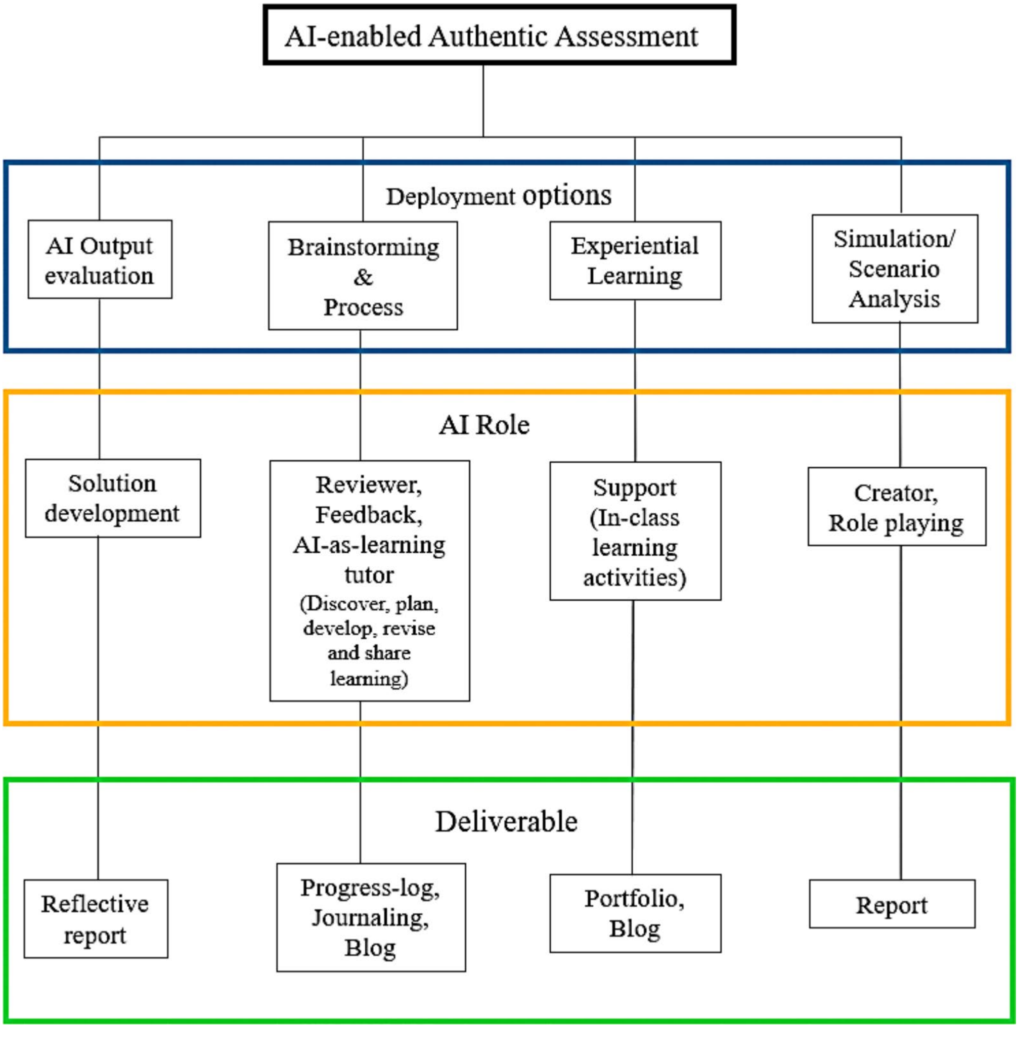

FIGURE 8 Proposed DAD conceptual framework for AI-enabled authentic assessment.

assessed. This agrees with the study of Kickbusch et al. (2025) that concluded assessment deliverable should evidence learning by situating them within a richer tapestry of process and reflection.

Limitations

The documents retrieved for the analysis are towards authentic assessment in higher education and not broad assessment. This implies that only assessment with focus on authenticity has been considered and thus, in terms of generalisation, academics, readers, educational policy makers, and education providers should take caution of this limitation. Secondly, in contrast to traditional methods like content analysis for synthesising literature, this study utilised a probabilistic topic model to uncover the hidden themes within the research documents. Thereafter, this paper reviewed empirical findings of the documents with high contribution within each topic. This implies that out of the 103 documents used for the analysis, findings from the documents with low topic contribution were not considered.

More importantly, the DAD conceptual framework does not account for learners' differences (for example, AI literacy levels) and, as such, users should provide adequate scaffolding support to cater to this.

CONCLUSION

This study set out to analyse AI-enabled authentic assessment strategies through evidence synthesis. We analysed 103 documents collected from seven databases using term co-occurrence analysis and TM. The results from the co-occurrence network map showed emerging themes are academic integrity, active learning, concern, critical thinking, framework, GenAI, implication, integration, and issue. Further to this, the network map produced five main clusters, namely, effectiveness and integration of AI (cluster 1), critical thinking and learning strategies (cluster 2), teaching, learning and technology (cluster 3), impact, case study and assessment question (cluster 4), issues, and academic integrity (cluster 5). While the TM results suggest themes such as evaluation, framework, and skill development, teaching and learning methods, review of practice, application of knowledge, course-specific assessment design, and academic integrity. The results of the qualitative synthesis of literature within the topics showed that frameworks (as given in Table 7) do exist to guide academics in the preparation of AI-enabled authentic assessments but are limited in the AI integration options and scope. Thus, this study proposes the DAD conceptual framework for AI-enabled authentic assessment (shown in Figure 8) to organise knowledge and guide the future design and empirical testing, aligning claims with evidence. To conclude, key recommendations from this study are summarised as follows:

- While ensuring learning outcomes are being achieved, academics should also ensure that assessment fulfils the core principles of authenticity (realism, cognitive challenge, and evaluative judgement) and the assessment task should be well aligned with the development of students' knowledge, skills, and behaviours.
- Academics should consider incorporating AI in multiple phases of the assessment deliverables for both formative and summative assessments.
- The assessment marking rubrics should place greater emphasis on assessing the criteria related to the cognitive skills process of the assessment deliverable
- To adopt the output evaluation, academics need to consider layers of assessment tasks during the assessment design.
- Academics need to undergo adequate training in understanding the characteristics and vulnerabilities of AI systems to promote responsible usage.

AUTHOR CONTRIBUTIONS

Olakunle Olayinka: Validation; writing – review and editing; writing – original draft; investigation; formal analysis; project administration; resources. **Bayode Ogunleye:** Conceptualization; investigation; writing – original draft; methodology; validation; visualization; writing – review and editing; software; formal analysis; project administration; supervision; data curation; resources; funding acquisition. **Hemlata Sharma:** Writing – review and editing; writing – original draft; project administration; formal analysis; investigation; resources. **Oluwaseun Ajao:** Writing – review and editing; writing – original draft; project administration; formal analysis; investigation; resources. **Kudirat Ibilola Zakariyyah:** Writing – original draft; writing – review and editing; formal analysis; project administration; investigation; resources.

ACKNOWLEDGEMENTS

We acknowledge that there are no grants and support received.

AI STATEMENT

Authors have declared no AI tools were used in the preparation of this work.

FUNDING INFORMATION

The authors received no financial support for the research, authorship, and/or publication of this article.

CONFLICT OF INTEREST STATEMENT

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ETHICS STATEMENT

The article contains no studies involving human participants performed by any author.

ORCID

Bayode Ogunleye https://orcid.org/0000-0001-6178-0731

REFERENCES

Abbasnejad, B., Soltani, S., Taghizadeh, F., & Zare, A. (2025). Developing a multilevel framework for AI integration in technical and engineering higher education: Insights from bibliometric analysis and ethnographic research. *Interactive Technology and Smart Education*, *23*, 49–79.

Acar, O. A. (2023). *Are your students ready for AI? A 4-step framework to prepare learners for a ChatGPT world*. Harvard Business Publishing Education. www.kcl.ac.uk/about/strategy/learning-and-teaching/ai-guidance/pair-framework-guidance

Alqarni, A. (2025). Artificial intelligence-critical pedagogic: Design and psychologic validation of a teacher-specific scale for enhancing critical thinking in classrooms. *Journal of Computer Assisted Learning*, *41*(3), e70039.

Alshwiah, A. (2024). Students' perceptions of an artificially intelligent chatbot as a support tool to develop their research skills. *Qualitative Research Journal*, *26*, 168–179.

Antoniou, C., Pavlou, A., & Ikossi, D. G. (2025). Let's chat! Integrating ChatGPT in medical student assignments to enhance critical analysis. *Medical Teacher*, *47*(5), 791–793.

Ashford-Rowe, K., Herrington, J., & Brown, C. (2014). Establishing the critical elements that determine authentic assessment. *Assessment & Evaluation in Higher Education*, *39*(2), 205–222.

Aydin, G., & Ogunleye, B. (2026). What consumers really think about sustainable fashion: A computational analysis of online discussions. *Journal of Global Fashion Marketing*, *24*, 1–31.

Bannister, P., Urbieta, A. S., & Alvira, N. B. (2025). Appraising higher education assessment validity: Development of the PANDORA GenAI susceptibility rubric. *Journal of Applied Learning and Teaching*, *8*(1), 41–55.

Bird, S. (2006). NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions* (pp. 69–72). Association for Computational Linguistics.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Boud, D. (2012). Assessment and learning: Contradictory or complementary? In *Assessment for learning in higher education* (pp. 35–48). Routledge.

Chen, Y., Chai, K. K., Loo, J., Moosaei, R., & Obstfeld, J. (2025). GenAI-empowered group-based authentic assessment for network engineering courses. In *2025 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1–7). IEEE.

Choi, J. H., Hickman, K. E., Monahan, A. B., & Schwarcz, D. (2021). ChatGPT goes to law school. *Journal of Legal Education*, *71*, 387.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Corbin, T., Bearman, M., Boud, D., & Dawson, P. (2025). The wicked problem of AI and assessment. *Assessment & Evaluation in Higher Education*, *51*, 1–17.

Davey, S. K., Birbeck, D., Nallaya, S., Sallows, G., & Della Vedova, C. B. (2025). Utilising one-on-one interactive oral assessments as the major final assessment within a bioscience course. *Assessment & Evaluation in Higher Education*, *50*, 1–18.

Donnelly, R., & Kennelly, I. (2024). A framework-based cross-institutional CPD for academic staff in gen-AI literacy, critical inquiry, and authentic assessment. In *Effective practices in AI literacy education: Case studies and reflections* (pp. 51–59). Emerald Publishing Limited.

Dwight, J. (2025). Designing and constructing authentic assessments for cybersecurity education. In *2025 13th International Conference on Information and Education Technology (ICIET)* (pp. 247–251). IEEE.

Elson, L., Gaughan, L., Hopton, B., Lloyd-Brown, S., Briggs, C., Gharamti, M., & Moriarty, P. (2025). ChatGPT did my homework': Lessons learnt from embedding a chatbot in undergraduate coursework. *Physics Education*, *60*(2), 025022.

Essien, A., Bukoye, O. T., O'Dea, X., & Kremantzis, M. (2024). The influence of AI text generators on critical thinking skills in UK business schools. *Studies in Higher Education*, *49*, 1–18.

Farrell, H. (2024). Re-imagining student engagement in an AI-enhanced classroom: Strategies and practices. In *Using generative AI effectively in higher education* (pp. 67–73). Routledge.

Fatima, S. S., Sheikh, N. A., & Osama, A. (2024). Authentic assessment in medical education: Exploring AI integration and student-as-partners collaboration. *Postgraduate Medical Journal*, *100*(1190), 959–967.

Foung, D., Lin, L., & Chen, J. (2024). Reinventing assessments with ChatGPT and other online tools: Opportunities for GenAI-empowered assessment practices. *Computers and Education: Artificial Intelligence*, *6*, 100250.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.

Gulikers, J. T., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, *52*(3), 67–86.

Hamel, P., & Lee, W. K. (2024). Supporting the evaluation of authentic assessment in environmental sciences: A case study. *Cogent Education*, *11*(1), 2399433.

Hatley, N., & Penny, P. (2024). Embracing GenAI in education: A path towards authentic assessment. In *Using generative AI effectively in higher education* (pp. 87–93). Routledge.

Hsiao, Y. P., Klijn, N., & Chiu, M. S. (2023). Developing a framework to re-design writing assignment assessment for the era of large language models. *Learning: Research and Practice*, *9*(2), 148–158.

Hughes, H. P., Padgett, R. C., & Donald, W. E. (2025). Preparing students for hybrid working: The place and case for authentic assessment via synchronous online presentations. *Industry and Higher Education*, *39*(2), 134–138.

Khan, M. M., Dong, Y., & Manesh, N. A. (2023). Authentic assessment design for meeting the challenges of generative artificial intelligence. In *2023 IEEE Frontiers in Education Conference (FIE)* (pp. 1–8). IEEE.

Khojasteh, L., Kafipour, R., Pakdel, F., & Mukundan, J. (2025). Empowering medical students with AI writing co-pilots: Design and validation of AI self-assessment toolkit. *BMC Medical Education*, *25*(1), 159.

Kickbusch, S., Ashford-Rowe, K., Kemp, A., Boreland, J., & Huijser, H. (2025). Beyond detection: Redesigning authentic assessment in an AI-mediated world. *Education Sciences*, *15*(11), 1537.

Kizhakkethil, P., & Perryman, C. (2024). Student perceptions of generative AI in LIS coursework. *Proceedings of the Association for Information Science and Technology*, *61*(1), 531–536.

Kofinas, A. K., Tsay, C. H. H., & Pike, D. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. *British Journal of Educational Technology*, *56*, 2522–2549.

Kunitsu, Y. (2023). The potential of GPT-4 as a support tool for pharmacists: Analytical study using the Japanese national examination for pharmacists. *JMIR Medical Education*, *9*, e48452.

Lee, D., Arnold, M., Srivastava, A., Plastow, K., Strelan, P., Ploeckl, F., Lekkas, D., Palmer, E., & Palmer, E. (2024). The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Computers and Education: Artificial Intelligence*, *6*, 100221.

Lehane, S., & Wright, A. (2024, June). Designing authentic assessment to improve academic integrity. In *10th international conference on higher education advances (HEAd'24)* (pp. 18–21). https://doi.org/10.4995/HEAd24.2024.17136

Lim, J., & Hwang, J. (2025). Exploring trends and topics in hybrid intelligence using keyword co-occurrence networks and topic modelling. *Futures*, *167*, 103550.

Lim, J. P., & Lauw, H. W. (2024). Aligning human and computational coherence evaluations. *Computational Linguistics*, *50*(3), 893–952.

Malinka, K., Peresíni, M., Firc, A., Hujňák, O., & Janus, F. (2023). On the educational impact of chatgpt: Is artificial intelligence ready to obtain a university degree? In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1* (pp. 47–53). Association for Computing Machinery.

Martin, A. F., Tubaltseva, S., Harrison, A., & Rubin, G. J. (2025). Participatory Co-design and evaluation of a novel approach to generative AI-integrated coursework assessment in higher education. *Behavioral Sciences*, *15*(6), 808.

McCauley, J., Glantz, E. J., Nasereddin, M., & Bartolacci, M. R. (2025, November). Undergraduate cybersecurity redesign for the GenAI era: Beyond authentic assessment. In *Proceedings of the 26th ACM Annual Conference on Cybersecurity & Information Technology Education* (pp. 47–51). Association for Computing Machinery.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Minagar, S., Sakzad, A., Tack, G., Rudolph, C., & Sheard, J. (2024). ALAN: Assessment-as-learning authentic tasks for networking. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* (pp. 853–859). Association for Computing Machinery.

Miserandino, M. (2025). Authentic and creative assessment in a world with AI. *Teaching of Psychology*, *52*(3), 267–272.

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, *5*, 100151.

Murphy, V., Fox, J., Freeman, S., & Hughes, N. (2017). "Keeping it Real": A review of the benefits, challenges and steps towards implementing authentic assessment. *All Ireland Journal of Higher Education*, *9*(3).

Nguyen, K. V. (2025). The use of generative AI tools in higher education: Ethical and pedagogical principles. *Journal of Academic Ethics*, *23*, 1–21.

Ogunleye, B., Lancho Barrantes, B. S., & Zakariyyah, K. I. (2025). Topic modelling through the bibliometrics lens and its technique. *Artificial Intelligence Review*, *58*(3), 74.

Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of topic modelling approaches in the banking context. *Applied Sciences*, *13*(2), 797.

Ogunleye, B., Zakariyyah, K. I., Ajao, O., Olayinka, O., & Sharma, H. (2024a). Higher education assessment practice in the era of generative AI tools. *Journal of Applied Learning and Teaching*, *7*(1), 46–56. https://doi.org/10.37074/jalt.2024.7.1.28

Ogunleye, B., Zakariyyah, K. I., Ajao, O., Olayinka, O., & Sharma, H. (2024b). A systematic review of generative AI for teaching and learning practice. *Education Sciences*, *14*(6), 636.

O'Riordan, F., Thangaraj, J., Girme, P., & Ward, M. (2025). Interactive oral assessment: Staff perceptions, challenges and benefits of this robust, authentic assessment design approach. *Innovations in Education and Teaching International*, *63*, 1–14.

Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers & Education: Artificial Intelligence*, *2*, 100020.

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. *BMJ*, *372*, n71.

Peters, M., & Angelov, D. (2025). Redefining assessment tasks to promote students' creativity and integrity in the age of generative artificial intelligence. *International Journal for Educational Integrity*, *21*(1), 25.

Ping, F., Kanesin, T., & Law, L. F. (2025). Integration of AI tutoring in assessment designs for adult learners. *SN Computer Science*, *6*(5), 1–10.

Powell, S., & Forsyth, R. (2024). Generative AI and the implications for authentic assessment. In *Using generative AI effectively in higher education* (pp. 97–105). Routledge.

Pyakurel, P., & Souppez, J. B. R. (2025). Embedding authenticity in assessments for engineering education. In *Conference Proceedings. New Perspectives in Science Education 2025*. Filodiritto Editore. https://www.filodiritto.com/proceedings

Qi, L., & Zhu, J. (2024). Visualizing research trends on the use of generative AI in assessment in the WOS database from 2019 to 2024 via Vosviewer and CiteSpace. In *Proceedings of the 2024 8th International Conference on E-Society, E-Education and E-Technology* (pp. 28–35). Association for Computing Machinery.

Ramakrishnan, C., Cassidy, S., & Bower, M. (2025). Evaluating student performance and interactions in generative AI-integrated SQL practical tests. In *Proceedings of the 30th ACM Conference on Innovation and Technology in Computer Science Education V. 2* (p. 786). Association for Computing Machinery.

Rasul, T., Nair, S., Kalendra, D., Balaji, M. S., de Oliveira Santini, F., Ladeira, W. J., Rather, R. A., Yasin, N., Rodriguez, R. V., Kokkalis, P., Murad, M. W., & Hossain, M. U. (2024). Enhancing academic integrity among students in GenAI Era: A holistic framework. *The International Journal of Management Education*, *22*(3), 101041.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*. University of Malta.

Reimer, E. C. (2024). Examining the role of generative AI in enhancing social work education: An analysis of curriculum and assessment design. *Social Sciences*, *13*(12), 648.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on web search and data mining* (pp. 399–408). Association for Computing Machinery.

Sakzad, A., Paul, D., Sheard, J., Brankovic, L., Skerritt, M. P., Li, N., Minagar, S., Simon, & Billingsley, W. (2024). Diverging assessments: What, why, and experiences. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* (pp. 1161–1167). Association for Computing Machinery.
Salinas-Navarro, D. E., Vilalta-Perdomo, E., Michel-Villarreal, R., & Montesinos, L. (2024a). Using generative artificial intelligence tools to explain and enhance experiential learning for authentic assessment. *Education Sciences*, *14*(1), 83.
Salinas-Navarro, D. E., Vilalta-Perdomo, E., Michel-Villarreal, R., & Montesinos, L. (2024b). Designing experiential learning activities with generative artificial intelligence tools for authentic assessment. *Interactive Technology and Smart Education*, *21*(4), 708–734.
Sinnappan, G. S., Shauki, N. B. I., Cheng, W. F., & Teh, L. W. (2025). MetaC framework: Metacognitive prompting for AI-driven coursework assessment in tertiary education. In *2025 17th International Conference on Knowledge and Smart Technology (KST)* (pp. 145–150). IEEE.
Su, F., O'Dea, M., Wood, M., & Seidl, E. (2026). Embracing a new world: Authentic assessment designs in an age of generative artificial intelligence (GenAI). *Perspectives: Policy and Practice in Higher Education*, 1–12. https://doi.org/10.1080/13603108.2025.2601741
Swaffield, S. (2011). Getting to the heart of authentic assessment for learning. *Assessment in Education: Principles, Policy & Practice*, *18*(4), 433–449.
Tertiary Education Quality and Standards Agency. (2023). Assessment reform in the age of artificial intelligence. https://www.teqsa.gov.au/sites/default/files/2023-09/assessment-reform-age-artificial-intelligence-discussion-paper.pdf
Thanh, B. N., Vo, D. T. H., Nhat, M. N., Pham, T. T. T., Trung, H. T., & Xuan, S. H. (2023). Race with the machines: Assessing the capability of generative AI in solving authentic assessments. *Australasian Journal of Educational Technology*, *39*(5), 59–81.
Tran, M., Balasooriya, C., Jonnagaddala, J., Leung, G. K. K., Mahboobani, N., Ramani, S., Rhee, J., Schuwirth, L., Najafzadeh-Tabrizi, N. S., Semmler, C., & Wong, Z. S. (2025). Situating governance and regulatory concerns for generative artificial intelligence and large language models in medical education. *Npj Digital Medicine*, *8*(1), 315.
Upsher, R., Heard, C., Yalcintas, S., Pearson, J., & Findon, J. (2024). Embracing generative AI in authentic assessment: Challenges, ethics, and opportunities. In *Using generative AI effectively in higher education* (pp. 106–114). Routledge.
Vallis, C. (2025). Authentic assessment in higher education: The spectre of lost futures. *Teaching in Higher Education*, *30*(3), 744–751.
Van Eck, N. J., & Waltman, L. (2017). Citation-based clustering of publications using CitNetExplorer and VOSviewer. *Scientometrics*, *111*(2), 1053–1070.
Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment & Evaluation in Higher Education*, *43*(5), 840–854.
Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education*, *21*(1), 40.
Yung, R., Moyle, B., & Yang, E. C. (2024). Mitigating the impact of ChatGPT: Cultivating professional identity through interactive oral authentic assessments. *Journal of Hospitality & Tourism Education*, *37*, 1–7.
Zhao, C. (2024). AI-assisted assessment in higher education: A systematic review. *Journal of Educational Technology and Innovation*, *6*(4). https://doi.org/10.61414/jeti.v6i4.209

How to cite this article: Ogunleye, B., Olayinka, O., Zakariyyah, K. I., Ajao, O., & Sharma, H. (2026). An analysis of AI-enabled authentic assessment strategy using text mining techniques. *Review of Education*, *14*, e70206. https://doi.org/10.1002/rev3.70206

APPENDIX A1

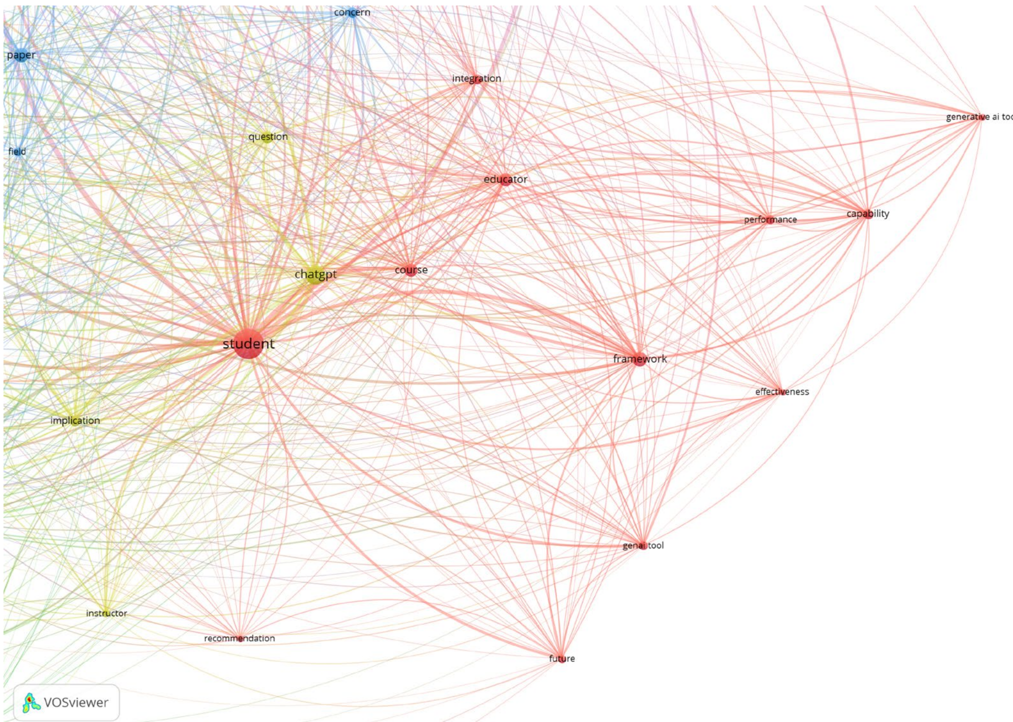

FIGURE 3A Cluster 1 (red).

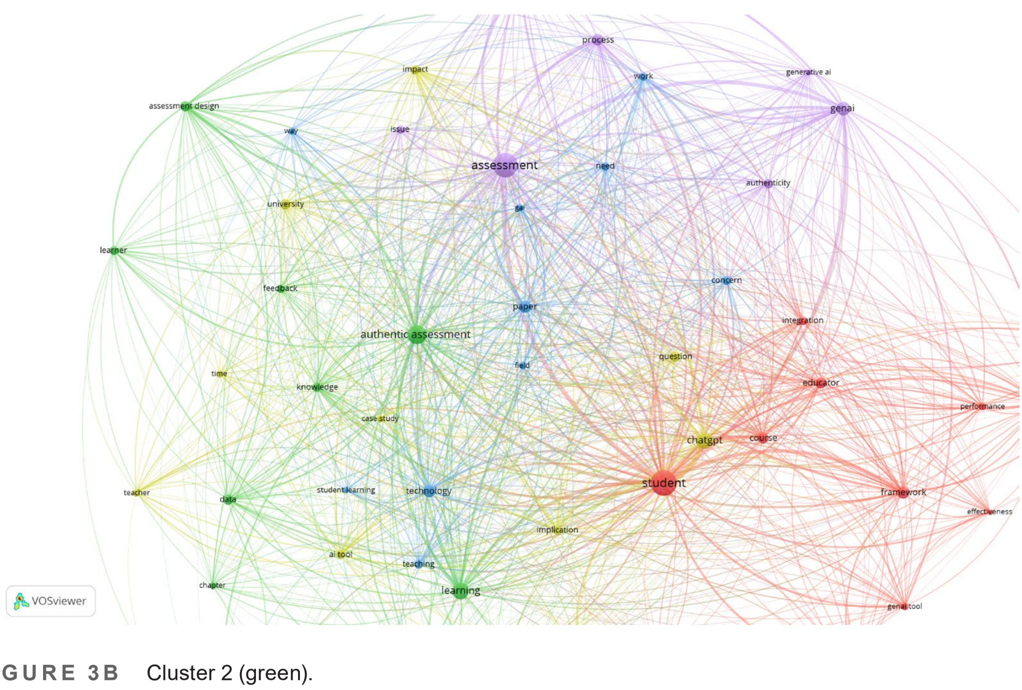

FIGURE 3B Cluster 2 (green).

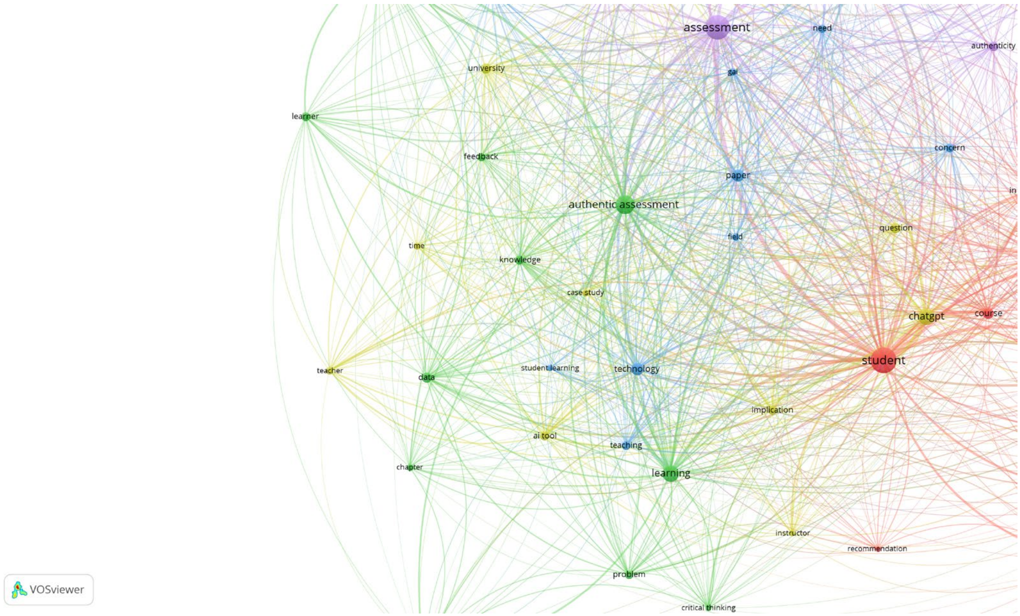

FIGURE 3C Cluster 3 (blue).

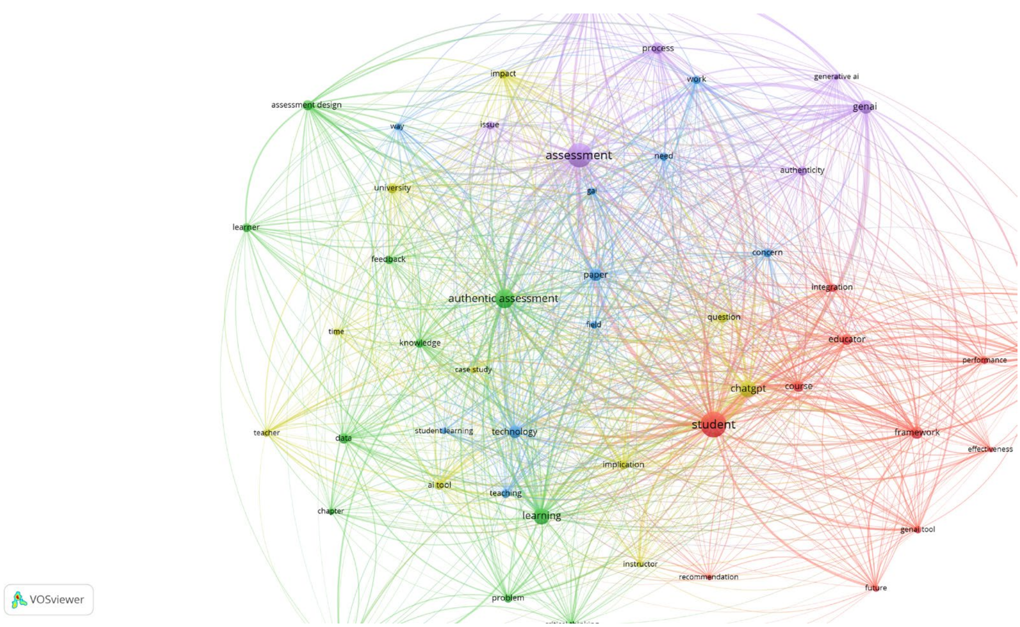

FIGURE 3D Cluster 4 (yellow).

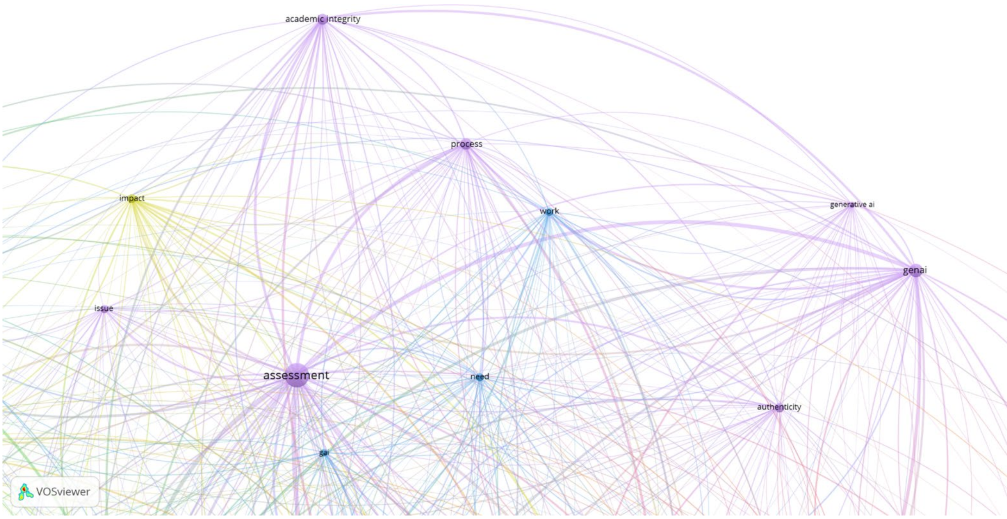

FIGURE 3E Cluster 5 (purple).

APPENDIX A2

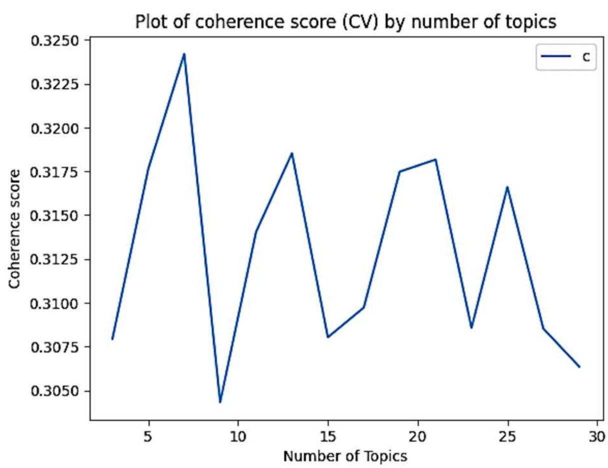

FIGURE 4 Plot of coherence score (CV) against number of topics.

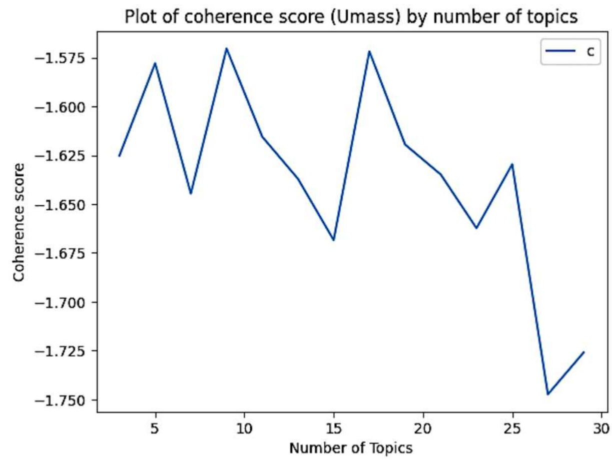

FIGURE 5 Plot of coherence score (UMass) against number of topics.

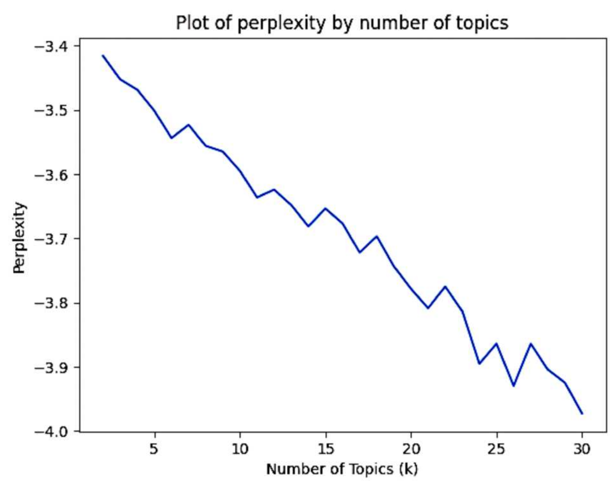

FIGURE 6 Plot of perplexity against number of topics.

```
Topic #1:
0.183*"evaluation" + 0.113*"challenge" + 0.096*"experience" + 0.076*"enhance" + 0.073*"skill" + 0.072*"opportunity" + 0.053*"process" + 0.047*"review" + 0.045*"practice" + 0.044*"thinking"

Topic #2:
0.184*"framework" + 0.153*"skill" + 0.078*"implication" + 0.077*"analysis" + 0.074*"challenge" + 0.060*"implementation" + 0.053*"development" + 0.049*"environment" + 0.048*"thinking" + 0.044*"course"

Topic #3:
0.346*"learning" + 0.101*"thinking" + 0.087*"concern" + 0.077*"opportunity" + 0.074*"experience" + 0.058*"integrity" + 0.054*"environment" + 0.041*"enhance" + 0.031*"application" + 0.028*"implication"

Topic #4:
0.136*"practice" + 0.106*"challenge" + 0.080*"task" + 0.078*"process" + 0.076*"implementation" + 0.075*"support" + 0.065*"application" + 0.065*"review" + 0.052*"environment" + 0.049*"literature"

Topic #5:
0.145*"application" + 0.099*"response" + 0.094*"challenge" + 0.093*"skill" + 0.084*"knowledge" + 0.074*"practice" + 0.058*"concern" + 0.046*"opportunity" + 0.045*"question" + 0.041*"course"

Topic #6:
0.151*"course" + 0.131*"question" + 0.107*"experience" + 0.074*"development" + 0.072*"process" + 0.057*"educator" + 0.055*"response" + 0.049*"practice" + 0.049*"result" + 0.039*"framework"

Topic #7:
0.256*"integrity" + 0.103*"educator" + 0.089*"practice" + 0.060*"challenge" + 0.059*"issue" + 0.053*"analysis" + 0.044*"opportunity" + 0.041*"concern" + 0.040*"development" + 0.035*"process"
```

FIGURE 7 LDA topics with weights.