Journal of Computer Education
Knowing is not enough, you must be able to do it.
Vol. 4(2), pp. 1-20, December 2025
https://www.journalofcomputereducation.info/

Beyond Recommendation: A Critical Review of Generative AI in Educational Recommender Systems

Edgar Ceh-Varela*
Department of Mathematical Sciences, Eastern New Mexico University, Portales, USA
ORCID: *0000-0001-6277-2741*

Yitzen Lizama-Peraza
Department of Languages and Literature, Eastern New Mexico University, Portales, USA
ORCID: 0009-0001-4726-2900

Article history
Received:
31.07.2025

Received in revised form:
01.10.2025

Accepted:
03.10.2025

Key words:
generative ai, large language models, educational technology, content generation, recommender systems

The emergence of Generative AI (GenAI) promises to shift educational recommender systems from static content curation to dynamic experience creation. But is this technical promise translating into demonstrable educational value? This paper presents a critical review, grounded in a comprehensive search of 1,223 articles across four major databases (ACM, IEEE, ScienceDirect, MDPI), which yielded only 16 foundational studies meeting our specific criteria, to argue that a significant "evaluation gap" is hindering the field's progress. This scarcity itself is a key finding, confirming the field's nascent state. Our analysis of this core literature reveals a field focused on technical innovation. Applications are dominated by Learning Path Recommendation (63%) and Content Generation (31%), driven by accessible techniques like Prompt Engineering (44%) and Fine-tuning (38%). These systems are successfully being reframed as generative co-pilots. However, we expose a critical misalignment: the methods used to evaluate these systems have not evolved with the technology. The literature is overwhelmingly reliant on technical and user-perception metrics (found in 82% of studies), with a near-total absence of research measuring direct improvements in student learning outcomes. This evaluation gap is the single most significant barrier to creating truly effective systems. This review's primary contribution is the evidence-based critique of this gap. We conclude by proposing a concrete research agenda focused on validating educational efficacy, ensuring pedagogical integrity, and building trustworthy systems to guide the field from its current state of technical potential toward a future of proven learning impact.

Introduction

Traditional recommender systems in education are tools designed to personalize learning experiences by suggesting relevant resources based on individual user preferences and behaviours. These systems play a crucial role in modern educational environments, where the

*Correspondency: eduardo.ceh@enmu.edu

sheer volume of available content can overwhelm learners. By utilizing techniques such as content-based filtering (Javed et al., 2021), collaborative filtering (Papadakis, Papagrigoriou, Panagiotakis, Kosmas, & Fragopoulou, 2022), and hybrid techniques (Seth & Sharaff, 2022), traditional recommender systems help tailor educational pathways, enhance user engagement, and ultimately improve academic outcomes across diverse educational contexts (Roy & Dutta, 2022; Ko, Lee, Park, & Choi, 2022).

However, these traditional systems are fundamentally limited to retrieving content from pre-existing catalogues (Harrathi & Braham, 2021). Their ability to deliver true one-to-one personalization is constrained, as they cannot generate new materials tailored to a learner's specific needs, interests, or real-time challenges. These systems can only suggest the closest available match rather than create ideal resources that address individual knowledge gaps.

The emergence of Generative AI (GenAI), particularly powerful Large Language Models (LLMs), promises to overcome this "content bottleneck" and drive a paradigm shift from static content recommendation to dynamic experience creation (Law, 2024). Tools such as ChatGPT (Welsby & Cheung, 2023), DALL-E (Marcus, Davis, & Aaronson, 2022), and Midjourney (Wahid, Mero, & Ritala, 2023) have democratized access to generative capabilities, enabling the creation of human-like text, realistic images, music, and more (Peñalvo & Ingelmo, 2023). However, this rapid technological progress raises a central, critical question: Is the promise of GenAI being met with the rigorous validation required to ensure it provides real educational value?

This paper presents a critical review that addresses this question directly. We argue that while the potential of GenAI in educational recommenders is immense, a significant "evaluation gap" between technical capability and pedagogical validation is hindering the field's progress. To ground this critique in evidence, we conducted a comprehensive search across four major databases (ACM, IEEE, ScienceDirect, and MDPI). Our search of over 1,200 articles yielded only 16 foundational peer-reviewed studies that met our specific inclusion criteria. This scarcity itself is a key finding, confirming that the core research in this area is highly nascent and allowing for a deep, focused analysis.

To structure our critical analysis of this foundational literature, we address the following four research questions:

- **RQ1:** What are the primary applications of GenAI in educational recommender systems?
- **RQ2:** What specific GenAI models and techniques are being employed?
- **RQ3:** How is the effectiveness of these GenAI-enhanced systems being evaluated?
- **RQ4:** What are the primary challenges and limitations identified in the literature?

By synthesizing the findings, this paper's primary contribution is to identify, characterize, and provide a roadmap to close the critical evaluation gap. We conclude by proposing a forward-looking agenda for developing effective, safe, and demonstrably valuable generative recommender systems for education.

The rest of this paper is organized as follows. The Literature Search and Scope Section describes the literature search strategy used to ground this review. The Results Section presents the results of our analysis. The Discussion Section discusses our main findings and their

implications. Finally, the Conclusions and Future Work Section presents our conclusions and a future work agenda.

Literature Search Strategy and Scope

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)[1] framework, which is the gold standard and has been used in similar studies (Barrera et al., 2023; Andrade-Ruiz et al., 2024; Valle-Cruz, Gil-Garcia, & Sandoval-Almazan, 2024; Law, 2024).

Database Selection

We selected ACM Digital Library[2], IEEE Xplore Digital Library[3], ScienceDirect[4], and MDPI[5] database as they represent the leading publishers of high-impact research in computer science, educational technology, and artificial intelligence.

Constructing the Search String

The goal for this step is to be comprehensive yet precise. Therefore, we built the string from our key concepts using Boolean operators (AND, OR). We defined three main concepts for our search string:

- **Concept 1: Generative AI**. We use the following keywords: “Generative AI”, “Generative Artificial Intelligence”, “Large Language Model”, “LLM”.
- **Concept 2: Recommender Systems.** We use the following keywords: “Recommender System”, “Recommendation System”.
- **Concept 3: Education.** We use the following keywords: “Education”, “Learning”, “Student”.

We combined these with AND between concepts and OR within concepts. We started with a more focused string and broadened it if we obtained too few results. For instance, we started with:

“Generative AI” AND “Recommender System” AND “Education”

The full Boolean search string used was:

("Generative AI” OR “Generative Artificial Intelligence” OR “Large Language Model *” OR “LLM *”)

AND (“Recommender System *” OR “Recommendation System *”)

AND (“Education *” OR “Learning” OR “Student *”)

[1] https://www.prisma-statement.org
[2] https://dl.acm.org
[3] https://ieeexplore.ieee.org
[4] https://www.sciencedirect.com
[5] https://www.mdpi.com

The exact syntax may vary slightly between databases. Truncation symbols (e.g., *) were used to capture variations of key terms We searched for these terms in the title, abstract, and author-specified keywords.

To maintain a tight focus on recommender systems as a core technology, the search query was constructed to be highly specific. Broader, related terms such as "adaptive learning," "intelligent tutoring," and "personalized learning" were deliberately excluded from the search string. While this approach may risk missing some relevant studies that do not use the term "recommender system," it significantly increases the precision of the search and ensures that the final corpus of papers is directly centered on the review's primary research questions.

Inclusion and Exclusion Criteria

The resulting articles were then screened for inclusion based on a predefined set of criteria. The screening was performed in two stages: first on the title and abstract, and second on the full text. The selection of final papers was guided by a precise set of inclusion criteria, outlined in Table 1. These rules were applied consistently during the screening and eligibility phases to ensure the final corpus of literature was directly relevant, academically rigorous, and representative of the current state-of-the-art.

The central criterion mandated that the paper's primary topic involved the use of GenAI for recommendation within an educational context, ensuring the exclusion of papers with only tangential mentions. For quality assurance, we limited our selection to peer-reviewed publications, thereby excluding editorials, pre-prints, and non-academic articles. The temporal scope was intentionally set from 2021 onwards to align with the emergence and widespread impact of large-scale foundation models on the field of AI. Lastly, for feasibility, the review was constrained to articles published in English and for which the full text could be obtained for detailed evaluation.

Table 1. Inclusion criteria for paper selection

Criteria	Inclusion
Topic	Paper's primary focus is the use of GenAI for recommendation in an educational setting
Publication Type	Peer-reviewed journal articles, conference proceedings, workshop papers
Language	English
Timeframe	Published from 2021 to Present
Access	Full text is accessible

Conversely, a paper was excluded if it met any of the exclusion criteria detailed in Table 2. These criteria were critical for refining the initial search results into a focused and high-quality corpus. The primary filter was thematic relevance; we systematically removed papers that, while potentially retrieved by our keywords, did not address the core intersection of our research questions. This included studies on legacy recommender systems, applications of GenAI in other domains like healthcare or finance, and general educational technology research without a significant GenAI component. To ensure the scientific validity of our synthesis, we also excluded all non-peer-reviewed materials.

Finally, practical and temporal constraints led to the exclusion of works published before 2021, those not written in English, and any articles for which full-text access could not be secured.

Table 2. Exclusion criteria for paper selection

Criteria	Exclusion
Topic	Papers on traditional recommender systems; papers on GenAI for non-educational tasks; papers on education without GenAI.
Publication Type	Editorials, opinions, pre-prints, keynotes, non-peer-reviewed articles
Language	Any other language
Timeframe	Published before 2021
Access	Abstract only

Study Selection

The study selection process followed the PRISMA 2020 framework to ensure a transparent and reproducible review, as illustrated in our flow diagram (Figure 1). The selection was conducted in two primary phases: (1) identification and screening of articles, and (2) assessment of full-text eligibility.

Phase 1: Identification and Screening

Our search strategy, conducted on June 6, 2025, was designed to be deliberately comprehensive, targeting four major academic databases to ensure broad coverage across computer science, educational technology, and multidisciplinary venues: ACM Digital Library, IEEE Xplore, ScienceDirect, and MDPI. The inclusion of ScienceDirect and MDPI was specifically to capture high-impact publications from leading journals that might lie outside the traditional ACM/IEEE ecosystems, such as Computers & Education: Artificial Intelligence, Education Sciences and AI.

Our search query yielded an initial set of 1,286 records: 904 from ACM, 216 from IEEE, 16 from ScienceDirect, and 150 from MDPI. Upon combining these results, 63 duplicate records were identified across the databases, indicating a distinct set of initial findings from each venue. Thus, 1,223 articles proceeded to the screening stage.

The first screening filter was applied to the publication type. A total of 401 records, from the ACM database, were excluded as they were not individual research articles but rather entries for entire conference proceedings, books, or editorials. This left 822 articles for a detailed title and abstract screening. In this second step, articles were evaluated for relevance against our strict inclusion criteria. A total of 801 articles were excluded because their primary focus was not on the application of GenAI within an educational recommender system framework.

A notable finding emerged from the screening process. Despite the breadth of ScienceDirect, a database that hosts numerous high-impact computer science and education journals, all 16 articles retrieved by our query were ultimately excluded. This was because, upon careful review, their primary focus did not align with our specific inclusion criteria for the application of GenAI within a recommender system framework.

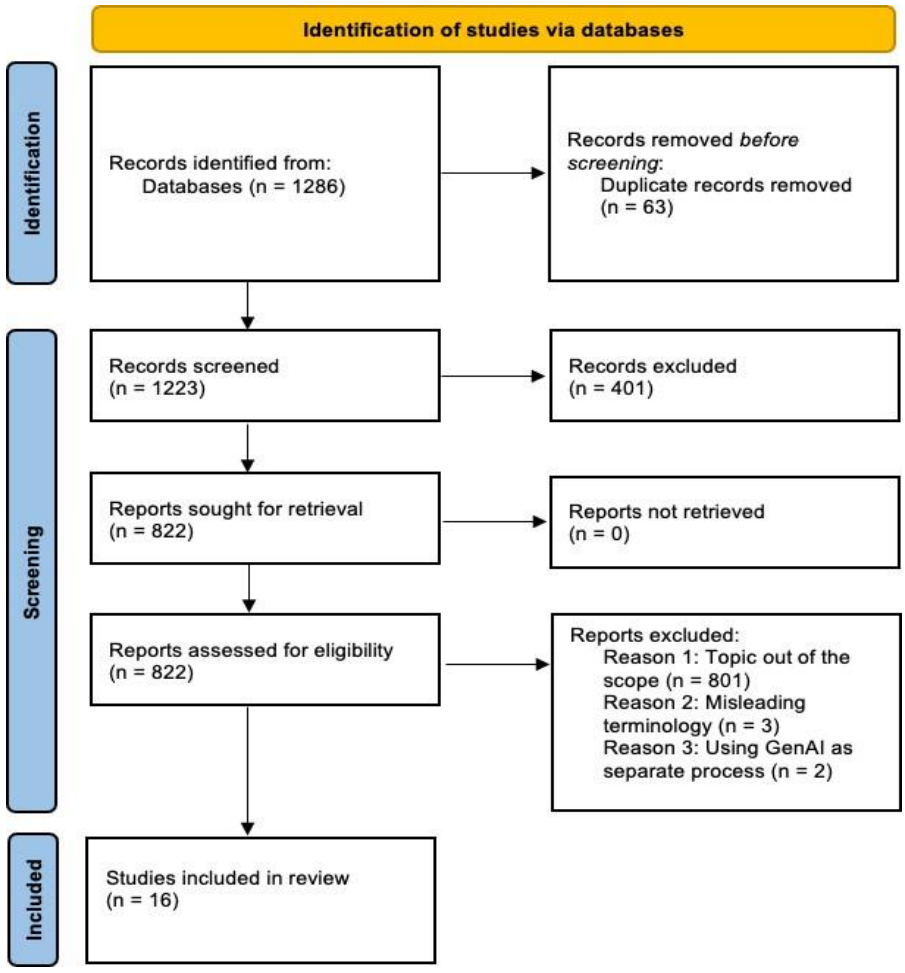

Figure 1. PRISMA flow diagram for the selection process for studies related to the systematic review.

Phase 2: Eligibility Assessment

This rigorous screening process resulted in 21 studies being selected for full-text retrieval and eligibility assessment. Upon detailed review of the full-text articles, five studies were discarded as they have a misleading terminology, or they use GenAI as a separate process. Only 16 were confirmed to meet the inclusion criteria and were therefore included in the final qualitative and quantitative synthesis.

Results

Overview of Included Studies

The systematic search and screening process concluded with a final corpus of 16 studies that met all inclusion criteria. All 16 articles selected after the title and abstract screening were successfully retrieved and confirmed for eligibility upon full-text review. These articles are presented in Table 3, which serves as a complete reference for the literature analyzed in this

review. For clarity and brevity in our analysis, each paper has been assigned a unique reference ID from [S1] to [S16], which will be used for citation purposes in the following sections.

Table 3. The 16 studies included in the systematic review.

ID	Author, year	Title
S1	Krouska et al., 2024 (Krouska, Troussas, Voyiatzis, Mylonas, & Sgouropoulou, 2024)	ChatGPT-based Recommendations for Personalized Content Creation and Instructional Design with a Tailored Prompt Generator
S2	Durrani et al., 2024 (Durrani et al., 2024)	Harnessing AI for Personalized Academic Major Recommendations: An Application of Large Language Models in Education
S3	Deepa et al., 2024 (Deepa, Rajkumar, & Balakrishnan, 2024)	Optimized Medical Recommendation System Utilizing Large Language Models for Enhanced Question Answering Performance
S4	Chun et al., 2024 (Chun, Ong, & Khong, 2024)	Reasonable Sense of Direction: Making Course Recommendations Understandable with LLMs
S5	Wang, 2025 (Wang, 2025)	Intelligent Recommendation of Open Education Teaching Resources based on Hybrid Collaborative Recommendation Algorithm with Large Language Model
S6	Hirnyi and Levus, 2024 (Hirnyi & Levus, 2024)	Educational and Competitive Algorithmic Platform Providing Personalized Recommendations
S7	Zhang et al., 2024 (Zhang, Zhao, & Zhou, 2024)	Application of Generative AI in the Teaching of Human Resource Management Course for Medical and Health Management Majors
S8	Abu-Rasheed et al., 2025 (Abu-Rasheed et al., 2025)	LLM-Assisted Knowledge Graph Completion for Curriculum and Domain Modelling in Personalized Higher Education Recommendations
S9	Chen et al., 2024 (Chen et al., 2024)	Course Recommendation System Based on Course Knowledge Graph Generated by Large Language Models
S10	Vishnumolakala et al., 2024 (Vishnumolakala, Sobin, Subheesh, Kumar, & Kumar, 2024)	AI-Based Research Companion (ARC): An Innovative Tool for Fostering Research Activities in Undergraduate Engineering Education
S11	Neyem et al., 2024 (Neyem et al., 2024)	Exploring the Impact of Generative AI for StandUp Report Recommendations in Software Capstone Project Development
S12	Frej et al., 2024 (Frej, Dai, Montariol, Bosselut, & Käser, 2024)	Course Recommender Systems Need to Consider the Job Market
S13	Wu et al., 2024 (Wu, Hu, Huang, Zeng, & Shao, 2024)	Intelligent Teaching Platform Based on Large Models
S14	Jamet et al., 2024 (Jamet, Manderlier, Shrestha, & Vlachos, 2024)	Evaluation and Simplification of Text Difficulty Using LLMs in the Context of Recommending Texts in French to Facilitate Language Learning
S15	Grimm and Rubart, 2024 (Grimm & Rubart, 2024)	Authoring Educational Hypercomics Assisted by Large Language Models
S16	Dahal et al., 2025(Dahal, Nugroho, Schmidt, & Sänger, 2025)	AI-Based Learning Recommendations: Use in Higher Education

This table serves as a foundational overview of the literature analyzed. Likewise, the distribution of these studies over time and by publication venue is presented in Table 4 and Figure 2. Notably, only three papers (18.75%) are from 2025. Similarly, the majority (62.5%) were published in IEEE venues. The key finding from the selection process itself is the high filtration rate, demonstrating the nascent state of this specific research domain.

Table 4. Publication venues of the included studies.

Publication Venue	Type	Database	Year	Count
International Conference on Foundation and Large Language Models (FLLM)	Conference	IEEE	2024	1
International Conference on Artificial Intelligence, Metaverse and Cybersecurity (ICAMAC)	Conference	IEEE	2024	1
Global Conference on Communications and Information Technologies (GCCIT)	Conference	IEEE	2024	1
International Midwest Symposium on Circuits and Systems (MWSCAS)	Conference	IEEE	2024	1
International Conference on Intelligent Systems and Computational Networks (ICISCN)	Conference	IEEE	2025	1
International Conference on Computer Science and Information Technologies (CSIT)	Conference	IEEE	2024	1
International Conference on Information Technology in Medicine and Education (ITME)	Conference	IEEE	2024	1
Global Engineering Education Conference (EDUCON)	Conference	IEEE	2025	1
International Conference on Teaching, Assessment and Learning for Engineering (TALE)	Conference	IEEE	2024	1
Global Engineering Education Conference (EDUCON)	Conference	IEEE	2024	1
Technical Symposium on Computer Science Education	Conference	ACM	2024	1
International ACM SIGIR Conference on Research and Development in Information Retrieval	Conference	ACM	2024	1
International Symposium on Artificial Intelligence for Education	Conference	ACM	2024	1
ACM Conference on Recommender Systems	Conference	ACM	2024	1
ACM Conference on Hypertext and Social Media	Conference	ACM	2024	1
Future Internet	Journal	MDPI	2025	1

Thematic Synthesis: Answering the Research Questions

Applications of GenAI (RQ1)

To address our first research question (RQ1) concerning the primary applications of Generative AI, we categorized the 16 included studies based on their main objective. The findings, summarized in Table 5, reveal two primary areas of focus: providing intelligent guidance on learning pathways and creating dynamic content and learning experiences.

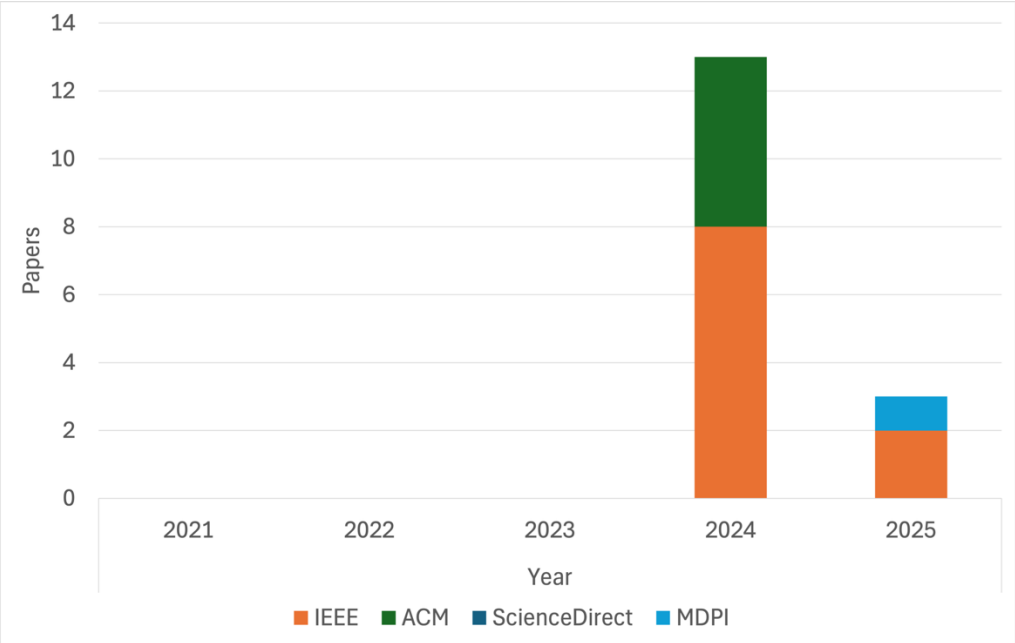

Figure 2. Number of studies per year

Table 5. Distribution of primary application areas across the 16 included studies.

Application Area	Count (n=16)	Percentage	References
Content Generation and Interaction	5	31%	S1, S3, S7, S13, S15
Learning Path Recommendation	10	63%	S2, S4, S5, S6, S8, S9, S10, S12, S14, S16
Assessment and Feedback	1	6%	S11
Total	**16**	**100%**	

The most prevalent application, accounting for more than half of the literature (n=10, 63%), was Learning Path Recommendation. These studies used GenAI to support academic decision-making, course selection, and the creation of personalized learning journeys. Some systems functioned as high-level academic advisors, helping students navigate major choices, research topics, and job-aligned course planning using advanced techniques like Retrieval-Augmented Generation and LLM-driven profiling [S2, S10, S12, S16]. Others focused on fine-grained course and content recommendation, combining collaborative filtering and knowledge graph modeling to personalize learning materials and make recommendations more transparent [S4, S5, S6, S9]. A final set of studies in this category emphasized adaptive learning journeys, generating personalized paths and language-specific reading materials based on learners' skills and progression needs [S8, S14].

The second most dominant application area (n=5, 31%) involves the use of GenAI for Content Generation and Interaction. This broader category includes studies that support both educators and learners through the creation of materials and direct, interactive assistance. On one hand, some systems focused on helping instructors design personalized lesson plans and creative materials like educational hypercomics [S1, S15]. On the other, a distinct subgroup delivered

dynamic, multimodal learning experiences through real-time AI tutoring and digital human teaching videos to enhance student engagement [S7, S13]. In a specialized context, one study demonstrated a model fine-tuned to generate accurate, context-aware responses for an expert-level domain such as medical education [S3].

Finally, a single study (n=1, 6%) focused specifically on Assessment and Feedback, suggesting this remains a nascent or challenging area of research within the literature. Overall, the distribution indicates a significant evolution from traditional recommenders. The field is increasingly focused on integrating GenAI not just to select, but to reason, create, and interact. Rather than solely generating content in isolation, the trend is to use GenAI to analyze learner data for intelligent guidance, and to provide dynamic experiences through tailored content and tutoring. This signals a decisive shift toward multifaceted, intelligent system that go far beyond static content recommendations.

GenAI Techniques Employed (RQ2)

To answer our second research question (RQ2) regarding the technologies employed, we analyzed the GenAI techniques and models described in the 16 studies. The findings, summarized in Table 6, reveal a strong reliance on large, pre-trained foundation models, with a clear preference for API-based access over more resource-intensive methods.

The 16 studies included in this review showcase a diverse yet strategically concentrated use of GenAI techniques in educational applications. As illustrated in Table 6, Prompt Engineering via API access was the most frequently employed technique, appearing in 7 out of 16 studies (44%). These implementations heavily leaned on ChatGPT, often without specifying the version [S1], although several studies did report using GPT-3.5 [S6, S11, S12, S15], GPT-4 [S6, S11], and even GPT-4o [S8] for more specialized tasks such as topic extraction or classification. Notably, multiple fine-tuned variants of LLaMA, including LongChat-7b-32k, LongChat-13b-16k, and Vicuna-33b-v1.3, were deployed via Chatbot Arena's API architecture, enabling scalable and customized AI-driven interactions [S4]. Other tools like Whisper, used for transcribing lecture videos [S8], further reflect the ecosystem-level integration of GenAI models.

Table 6. Distribution of primary GenAI techniques employed across the 16 included studies.

GenAI Technique	Count (n=16)	Percentage	References
Prompt Engineering (API)	7	44%	S1, S4, S6, S8, S11, S12, S15
Fine Tuning	6	38%	S3, S5, S9, S10, S13, S14
Retrieval Augmented Generation (RAG)	2	12%	S2 , S16
Not disclosed	1	6%	S7
Total	**16**	**100%**	

Fine-tuning emerged as the second most common technique, used in 6 studies (38%). These efforts often involved tailoring base models to better align with domain-specific educational tasks. A standout among these was GPT-3.5 and GPT-4, prominently featured for its fine-tuned performance in context-aware recommendations [S3, S5, S10]. Beyond OpenAI's models,

several studies turned to alternative LLMs such as ChatGLM-6B for automatic data generation [S9], ChatGLM2-6B for supervised fine-tuning [S13], and LlaMa [S9] for extracting prerequisite courses. In addition, RoBERTa-v3 was utilized to derive course embeddings [S9], and models like Mistral-7B and CamemBERT were adapted to support multilingual educational needs [S14], including difficulty estimation and text simplification.

Two studies [S2, S16] (12%) employed Retrieval-Augmented Generation (RAG) (Gao et al., 2023), combining Few-Shot Learning (FSL) (Song, Wang, Cai, Mondal, & Sahoo, 2023) with RAG techniques. While only one study specifies using OpenAI's ChatGPT-4o, the approach highlights a shift toward integrating structured knowledge retrieval into generative outputs to improve educational relevance and factuality.

Lastly, in one study [S7] (6%), the GenAI methodology was not disclosed. The authors referred only generally to the use of generative AI, offering no specific model names or architectural details.

In sum, while the distribution of GenAI techniques reflects a strong reliance on prompt engineering and fine-tuning, the specific models and infrastructures, ranging from OpenAI's GPT variants to open-source alternatives like LLaMA and ChatGLM, reveal an ecosystem that is both diverse and rapidly evolving.

Evaluation Methods and Metrics (RQ3)

Our third research question (RQ3) investigated the evaluation methodologies employed to assess the effectiveness of these GenAI-powered systems. As shown in Table 7 and Table 8, the current literature predominantly emphasizes system-level and user-centered evaluations, with a clear gap in the measurement of actual learning outcomes.

Table 7. Distribution of primary evaluation methods employed across the 16 included studies.

Evaluation Method	Count (n=16)	Percentage	References
Quantitative/Offline Analysis	8	50%	S1, S2, S3, S5, S7, S9, S12, S14
Expert Evaluation	2	12%	S4, S8,
User Study (Human Participants)	5	32%	S6, S10, S11, S15, S16
No Empirical Evaluation	1	6%	S13
Total	**16**	**100%**	

The most prevalent evaluation method was Quantitative/Offline Analysis, used in 50% (n=8) of the studies. These studies primarily assessed system performance using automated metrics such as Accuracy, Precision, Recall, F1-Score, BLEU, ROUGE, and ranking metrics like HR@10, MRR@10, and NDCG@10 [S1, S2, S3, S5, S7, S9, S12, S14]. This reflects a strong initial emphasis on algorithmic effectiveness and technical benchmarking. User Studies involving human participants were the next most common approach at 32% (n=5). These studies focused on user perception and satisfaction, evaluating constructs like usability, relevance of recommendations, motivation, and overall satisfaction through surveys and structured feedback [S6, S10, S11, S15, S16].

Expert Evaluation accounted for 12% (n=2), where domain experts assessed system outputs based on dimensions such as reasoning quality, hallucination mitigation, and usability [S4, S8]. These evaluations provided crucial qualitative insight into the pedagogical and factual soundness of the generated content.

Table 8. Common categories of evaluation metrics.

Metric Category	Specific Examples Found in Literature	References
System Performance	Accuracy, Precision, Recall, F1-Score, BLEU, ROUGE, HR@10, MRR@10, NDCG@10	S1, S2, S3, S5, S9, S12, S14
Content Quality (Expert-Rated)	Acceptance Rate (AR), Reasoning Quality, Hallucination Mitigation, Expert Feedback (usability, correctness)	S4, S8
User Perception and Satisfaction	User Satisfaction, Usability, Relevance / Accuracy of Recommendations, Perceived Usefulness / Effectiveness / Motivation	S6, S10, S11, S15, S16
Learning Gain	Test Scores, Course Completion Rate, Video Viewing Hours, Assignment Scores, Resource Utilization Rate	S7
No Empirical Evaluation	Learning Preferences, Course Progress, Activity Levels, Knowledge Mastery, Video Synchronization Accuracy	S13

Notably, only one study (6%) attempted to directly measure Learning Gain, employing outcome-based metrics like test scores, course completion, or assignment performance [S7]. Another study [S13] did not include any empirical evaluation, instead presenting a conceptual framework.

This distribution indicates that while the field demonstrates maturity in evaluating system functionality and user receptivity, there remains a significant gap in the empirical evidence connecting these GenAI tools to actual improvements in student learning. Future work must move beyond technical and perceptual metrics to prioritize the integration of direct learning assessments, such as pre/post-testing and long-term knowledge retention studies, to validate the educational efficacy of these powerful new systems.

Identified Challenges (RQ4)

Our final research question (RQ4) focused on identifying the challenges and limitations reported by researchers. A thematic analysis was conducted across all 16 papers, and the frequency of each challenge category is summarized in Table 9. The percentages reflect the proportion of unique studies that discussed each theme.

The most frequently cited concerns were Prompt Engineering and Interpretability and Dataset and Input Constraints, each mentioned in 50% of the studies (n=8). For prompt engineering, researchers pointed to the difficulty of crafting effective prompts and the opaque nature of LLM decision-making, which limits trust and understanding [S1, S2, S3, S4, S8, S14, S15, S16]. For dataset constraints, common issues included small and sparse datasets [S1, S5, S6, S12, S16], the underutilization of multimodal inputs [S1, S8], and a high dependency on manual validation of AI outputs [S9, S14].

Table 9: Frequency of challenges and limitations cited in the included studies. Note that studies could mention multiple challenges

Challenge / Limitation Category	Number of Studies Citing (out of 16)	Percentage of Studies Mentioning	References
Factual Accuracy and Hallucinations	4	25%	S3, S11, S14, S16
Pedagogical Soundness	5	31%	S1, S5, S7, S11, S13
Evaluation Difficulties	7	44%	S1, S4, S6, S12, S14, S15, S16
Bias and Fairness	5	31%	S2, S3, S10, S14, S15
Cost and Scalability	6	37%	S1, S4, S5, S6, S13, S16
Data Privacy and Security	6	37%	S2, S3, S4, S10, S11, S16
Prompt Engineering and Interpretability	8	50%	S1, S2, S3, S4, S8, S14, S15, S16
Dataset and Input Constraints	8	50%	S1, S5, S6, S8, S9, S12, S14, S16
Domain Adaptability and Transferability	4	25%	S2, S3, S10, S14

Evaluation Difficulties were highlighted in 44% of the studies (n=7), reinforcing our earlier findings from RQ3. Authors reported constraints such as small sample sizes [S1, S4, S6, S16] and the absence of real-world learning outcome assessments [S12, S14, S15], limiting the ability to robustly evaluate system effectiveness.

A significant cluster of challenges, each raised by more than a third of the literature, included Pedagogical Soundness, Bias and Fairness, Cost and Scalability, and Data Privacy and Security. Authors noted that GenAI outputs, while fluent, may fail to support meaningful learning by offering answers too directly or omitting scaffolding [S1, S5, S7, S11, S13]. They also cited concerns with propagating social biases [S2, S3, S10, S14, S15], high computational costs [S1, S4, S5, S6, S13, S16], and the need to protect sensitive learner data when using third-party APIs [S2, S3, S4, S10, S11].

Finally, Domain Adaptability and Transferability was discussed in 25% of studies (n=4), highlighting difficulties in adapting systems to niche domains [S2, S3, S10, S14]. Likewise, the issue of Factual Accuracy and Hallucinations was raised in 25% of studies (n=4), where researchers reported instances of fabricated content that could compromise learner trust [S3, S11, S14, S16].

Taken together, these self-reported limitations paint a picture of a field in transition, from proof-of-concept innovations to deployable, trustworthy learning technologies. While GenAI holds significant potential for transforming recommender systems for education, addressing these technical, pedagogical, and ethical challenges is essential to ensure that such systems are both effective and safe for learners.

Discussion

The findings of this systematic review confirm that the integration of Generative AI represents a fundamental paradigm shift within educational recommender systems. Our

analysis of the 16 included studies demonstrates a clear and accelerating transition away from the traditional model of static content curation and toward a new paradigm of dynamic experience creation. While the potential of this shift is profound, our findings also reveal a significant gap between technological capability and the pedagogical and evaluative rigor required for responsible deployment.

From Selecting Items to Creating Learning Experiences

Our analysis reveals a clear shift in the role of educational recommender systems, from curating existing content to actively generating new, personalized learning experiences. This evolution is most evident in the dominant application areas identified (RQ1). Learning Path Recommendation (63%) and Content Generation (31%) now account for the majority of use cases, marking a move away from simply selecting the best resources from a fixed repository. Instead, generative AI is increasingly being used to construct personalized learning paths, create custom instructional materials on demand, and provide interactive tutoring.

This transition is enabled by the accessibility of the underlying technology (RQ2). The growing use of API-based prompt engineering (44%) empowers researchers to focus on application design rather than model training, significantly lowering the barrier to entry for adding powerful generative capabilities to educational tools.

Taken together, these trends point to a redefinition of educational recommender systems. Rather than acting as passive selectors, like digital librarians, they are now evolving into generative co-pilots. These systems actively create content, provide guidance, and support personalized learning experiences. This evolution mirrors broader discussions in the literature, which highlight GenAI's potential to transform educational tools from static information providers into dynamic, interactive partners for learning (Meli, Taouki, & Pantazatos, 2024; Perifanou & Economides, 2025).

Aligning Technical Advances with Educational Goals

While researchers are actively exploring the creative potential of generative AI, our review shows a clear gap in how these systems are being evaluated (RQ3). The most striking finding is the lack of studies measuring real learning outcomes. Only one study [S7] directly assessed Learning Gain. Instead, most evaluations rely on Quantitative/Offline Analysis (50%), which focuses on system accuracy, and User Studies (32%) that measure user satisfaction or perception. While these are useful, they do not answer the most important question: do these educational recommender systems actually help students learn?

This gap is reflected in the challenges researchers report (RQ4). Many of the most common concerns go beyond technical performance. Issues such as Prompt Engineering and Interpretability (50%), Evaluation Difficulties (44%), Pedagogical Soundness (31%), and Factual Accuracy (25%) show that the field is struggling with a deeper problem. We are building systems that generate content easily and fluently, but we still lack strong methods to ensure that what they generate is pedagogically accurate, effective, and fair for learners. This challenge is not unique to GenAI. The broader field of educational technology has long struggled to move beyond engagement and satisfaction metrics to empirically demonstrate significant learning gains (Schindler, Burkholder, Morad, & Marsh, 2017; Bond & Bedenlier, 2019; Mallik & Gangopadhyay, 2023).

Making GenAI Recommender Systems Effective for Learning

The findings of this review offer important guidance for both researchers and practitioners in the field of generative AI in educational recommender systems. For researchers, the focus must shift from "Can we build it?" to "How do we prove they improve learning?" The gaps identified in this review suggest three key priorities:

- Develop evaluation methods that directly measure learning outcomes and long-term retention to validate true educational efficacy.
- Collaborate closely with educators and learning scientists to design systems that are pedagogically sound and learner-centered.
- Address technical challenges identified in this review, such as bias mitigation, factual accuracy, and privacy protection, within the context of generative recommenders.

For practitioners, including educators and developers, there is cause for cautious optimism. Generative AI-powered recommenders offer new possibilities for personalized learning pathways and content creation at scale. However, adoption should be guided by evidence of educational impact, not just user satisfaction or engagement metrics. Educators' roles will evolve from content providers to curators and overseers, ensuring these systems support meaningful and ethical learning experiences.

Limitations of the Review

It is important to interpret our findings within the methodological boundaries of this review. Several limitations should be noted.

First, the scope of our review was intentionally narrowed by the search terms used. By focusing on "recommender system" and "recommendation system," we aimed to prioritize analytical depth over broader topical coverage. As a result, relevant work in adjacent areas, such as intelligent tutoring systems or adaptive learning platforms that do not use these specific terms, may have been excluded.

Second, our database selection, while robust, is not exhaustive. Our search protocol included the ACM Digital Library, IEEE Xplore, ScienceDirect, and MDPI to ensure comprehensive coverage across both core computer science and leading educational technology venues. Notably, our rigorous screening found no articles from the ScienceDirect database that met our review's specific scope. However, our search is not exhaustive. Relevant studies could still reside in other databases or in niche journals not covered by our selection, and this remains a limitation of the review.

Third, we excluded pre-prints and restricted our review to peer-reviewed, English-language articles. This decision may have limited the inclusion of the most recent research and work from non-English-speaking scholarly communities.

Despite these constraints, this review offers a rigorous and focused snapshot of current research at the intersection of generative AI and educational recommender systems. It lays a foundation for future studies that may expand the scope, sources, and diversity of perspectives included.

Conclusions and Future Work

Generative AI is not merely enhancing educational recommender systems; it is forging a new identity for them. This systematic review of 16 peer-reviewed articles provides a comprehensive analysis of the emerging role of Generative AI in educational recommender systems. Our findings reveal a field undergoing a profound paradigm shift from passive content recommendation to the active creation of personalized learning paths, content, and interactions. The dominant technological approach involves leveraging powerful, pre-existing models through prompt engineering, enabling rapid innovation that transforms these recommender systems into generative co-pilots for learning.

However, our analysis reveals a significant misalignment between innovation and assessment. While the technical paradigm has advanced, the evaluative framework has not kept pace. We found a near-total absence of studies that empirically measure improvements in student learning outcomes. This "evaluation gap" represents the single greatest challenge facing the field. The very researchers driving this innovation consistently cite difficulties with evaluation, pedagogical soundness, and interpretability, confirming that technological enthusiasm has outpaced the methods required to validate true educational efficacy.

Bridging this gap is the defining work for the future of this field. We propose a research agenda focused on three critical pillars:

(1) Validating educational efficacy: Move beyond technical and perceptual metrics to prioritize robust, outcome-based evaluations. This requires adopting pre- and post-testing, long-term knowledge retention studies, and other methods from the learning sciences to prove that these systems actually help students learn.
(2) Ensuring pedagogical integrity: Foster deep collaboration between AI researchers and educational experts. The goal must be to embed sound learning principles directly into system design, ensuring that generative content is not just fluent but also effective, adaptive, and responsible.
(3) Building trustworthy systems: Prioritize solving the core technical challenges of accuracy, bias, and privacy. A system that hallucinates, perpetuates bias, or compromises student data cannot be considered a successful educational tool, no matter how engaging it is.

By moving beyond technical proof-of-concept to demonstrable educational value, the research community can harness the transformative potential of Generative AI to create truly effective and equitable learning experiences for all.

References

Abu-Rasheed, H., Jumbo, C., Al Amin, R., Weber, C., Wiese, V., Obermaisser, R., & Fathi, M. (2025). LLM-assisted knowledge graph completion for curriculum and domain modelling in personalized higher education recommendations. In *2025 IEEE Global Engineering Education Conference (EDUCON)* 1–5). IEEE. https://doi.org/10.48550/arXiv.2501.12300

Andrade-Ruiz, G., Carrasco, R.-A., Porcel, C., Serrano-Guerrero, J., Mata, F., &

Arias-Oliva, M. (2024). Emerging perspectives on the application of recommender systems in smart cities. *Electronics*, *13* (7), 1249. https://doi.org/10.3390/electronics13071249

Barrera, F. J., Brown, E. D., Rojo, A., Obeso, J., Plata, H., Lincango, E. P., . . . Shekhar, S. (2023). Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: A systematic review. *Frontiers in Endocrinology*, 14 , 1106625. https://doi.org/10.3389/fendo.2023.1106625

Bond, M., & Bedenlier, S. (2019). Facilitating student engagement through educational technology: Towards a conceptual framework. *Journal of Interactive Media in Education*, 2019 (1)k, 11. https://doi.org/10.5334/jime.528

Chen, X., Yin, C., Chen, H., Rong, W., Ouyang, Y., & Chai, Y. (2024). Course recommendation system based on course knowledge graph generated by large language models. In *2024 IEEE International Conference on Teaching, Assessment and Learning for Wngineering (TALE)* (pp. 1–8). IEEE. https://doi.org/10.1109/TALE62452.2024.10834324

Chun, H. W., Ong, R. K., & Khong, A. W. (2024). Reasonable sense of direction: Making course recommendations understandable with LLMs. *2024 IEEE 67th International Midwest Symposium on Circuits and Systems (MWSCAS),* 1408–1412. IEEE. https://doi.org/10.1109/MWSCAS60917.2024.10658914

Dahal, P., Nugroho, S., Schmidt, C., & Sänger, V. (2025). AI-based learning recommendations: Use in higher education. *Future Internet*, *17* (7), 285. https://doi.org/10.3390/fi17070285

Deepa, D., Rajkumar, T. D., & Balakrishnan, D. (2024). Optimized medical recommendation system utilizing large language models for enhanced question answering performance. *2024 Global Conference on Communications and Information Technologies (GCCIT),* 1–5. IEEE. https://doi.org/10.1109/GCCIT63234.2024.10862477

Durrani, U., Akpinar, M., Togher, M., Malik, A., Dordevic, M., & Aoudi, S. (2024). Harnessing AI for personalized academic major recommendations An application of large language models in education. *2024 International Conference on Artificial Intelligence, Metaverse and Cybersecurity (ICAMAC)*, 1–6. https://doi.org/10.1109/ICAMAC62387.2024.10828756

Frej, J., Dai, A., Montariol, S., Bosselut, A., & Käser, T. (2024). Course recommender systems need to consider the job market. In *Proceedings of the 47th international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 522–532). ACM. https://doi.org/10.1145/3626772.3657847

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., … Wang, H. (2023).

Retrieval-augmented generation for large language models: A survey. *arXiv.* https://doi.org/10.48550/arXiv.2312.10997

Grimm, V., & Rubart, J. (2024). Authoring educational hypercomics assisted by large language models. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media* (pp. 88–97). ACM. https://doi.org/10.1145/3648188.3675124

Harrathi, M., & Braham, R. (2021). Recommenders in improving students' engagement in large scale open learning. *Procedia Computer Science*, *192*, 1121–1131. https://doi.org/10.1016/j.procs.2021.08.115

Hirnyi, M., & Levus, Y. (2024). Educational and competitive algorithmic platform providing personalized recommendations. *2024 IEEE 19th International Conference on Computer Science and Information Technologies (CSIT),* 1–4. IEEE. https://doi.org/10.1109/CSIT65290.2024.10982648

Jamet, H., Manderlier, M., Shrestha, Y. R., & Vlachos, M. (2024). Evaluation and simplification of text difficulty using LLMs in the context of recommending texts in French to facilitate language learning. In *Proceedings of the 18th ACM Conference on Recommender Systems* (pp. 987–992). ACM. https://doi.org/10.1145/3640457.3688181

Javed, U., Shaukat, K., Hameed, I. A., Iqbal, F., Alam, T. M., & Luo, S. (2021). A review of content-based and context-based recommendation systems. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(3), 274–306. https://doi.org/10.3991/ijet.v16i03.18851

Ko, H., Lee, S., Park, Y., & Choi, A. (2022). A survey of recommendation systems: Recommendation models, techniques, and application fields. *Electronics*, *11*(1), 141. https://doi.org/10.3390/electronics11010141

Krouska, A., Troussas, C., Voyiatzis, I., Mylonas, P., & Sgouropoulou, C. (2024). ChatGPT-based recommendations for personalized content creation and instructional design with a tailored prompt generator. *In 2024 2nd International Conference on Foundation and Large Language Models (FLLM)* (pp. 295–299). IEEE. https://doi.org/10.1109/FLLM63129.2024.10852487

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*, 100174. https://doi.org/10.1016/j.caeo.2024.100174

Mallik, S., & Gangopadhyay, A. (2023). Proactive and reactive engagement of artificial intelligence methods for education: A review. *Frontiers in Artificial Iintelligence*, 6 ,1151391. https://doi.org/10.3389/frai.2023.1151391

Marcus, G., Davis, E., & Aaronson, S. (2022). A very preliminary analysis of DALL-E 2. *arXiv*. https://doi.org/10.48550/arXiv.2204.13807

Meli, K., Taouki, J., & Pantazatos, D. (2024). Empowering educators with generative AI: The GenAI Education Frontier Initiative. *In EDULEARN24 Proceedings* (pp. 4289–4299). https://doi.org/10.21125/edulearn.2024.1077

Neyem, A., Sandoval Alcocer, J. P., Mendoza, M., Centellas-Claros, L., Gonzalez, L. A., & Paredes-Robles, C. (2024). Exploring the impact of generative AI for standup report recommendations in software capstone project development. *In Proceedings of the 55th ACM Technical Symposium on Computer Science Education*, Vol. 1 (pp. 951–957). ACM. https://doi.org/10.1145/3626252.3630854

Papadakis, H., Papagrigoriou, A., Panagiotakis, C., Kosmas, E., & Fragopoulou, P. (2022). Collaborative filtering recommender systems taxonomy. *Knowledge and*

Information Systems, *64* (1), 35–74. ACM. https://doi.org/10.1007/s10115-021-01628-7

Peñalvo, F. J. G., & Ingelmo, A. V. (2023). What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in generative AI. *IJIMAI*, *8* (4), 7–16. https://doi.org/10.9781/ijimai.2023.07.006

Perifanou, M., & Economides, A. A. (2025). Collaborative uses of GenAI tools in project-based learning. *Education Sciences*, 15 (3), 354. https://doi.org/10.3390/educsci15030354

Roy, D., & Dutta, M. (2022). A systematic review and research perspective on recommender systems. *Journal of Big Data*, *9* (1), 59. https://doi.org/10.1186/s40537-022-00592-5

Schindler, L. A., Burkholder, G. J., Morad, O. A., & Marsh, C. (2017). Computer-based technology and student engagement: A critical review of the literature. *International Journal of Educational Technology in Higher Education, 14* (1), 25. https://doi.org/10.1186/s41239-017-0063-0

Seth, R., & Sharaff, A. (2022). A comparative overview of hybrid recommender systems: Review, challenges, and prospects. *Data Mining and Machine Learning Applications*, 57–98. https://doi.org/10.1002/9781119792529.ch3

Song, Y., Wang, T., Cai, P., Mondal, S. K., & Sahoo, J. P. (2023). A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, 55 (13s), 1–40. https://doi.org/10.1145/3582688

Valle-Cruz, D., Gil-Garcia, J. R., & Sandoval-Almazan, R. (2024). Artificial intelligence algorithms and applications in the public sector: A systematic literature review based on the PRISMA approach. In *Research Handbook on Public Management and Artificial Intelligence*, 8–26. Elgar. https://doi.org/10.4337/9781802207347.00010

Vishnumolakala, S. K., Sobin, C., Subheesh, N., Kumar, P., & Kumar, R. (2024). AI-based Research Companion (ARC): An innovative tool for fostering research activities in undergraduate engineering education. In 2024 IEEE Global Engineering Education Conference (EDUCON) (pp. 1–5). IEEE. https://doi.org/10.1109/EDUCON60312.2024.10578646

Wahid, R., Mero, J., & Ritala, P. (2023). Written by ChatGPT, illustrated by MidJourney: Generative AI for content marketing. *Asia Pacific Journal of Marketing and Logistics*, 35 (8), 1813–1822. https://doi.org/10.1108/APJML-10-2023-994

Wang, S. (2025). Intelligent recommendation of open education teaching resources based on hybrid collaborative recommendation algorithm with large language model. In *2025 International Conference on Intelligent Systems and Computational Networks (ICISCN)* (pp. 1–9). https://doi.org/10.1109/ICISCN64258.2025.10934397

Welsby, P., & Cheung, B. M. (2023). ChatGPT. *Postgraduate Medical Journal*, 99(1176), 1047-1048. Oxford University Press. https://doi.org/10.1093/postmj/qgad056

Wu, W., Hu, J., Huang, Y., Zeng, W., & Shao, H. (2024). Intelligent teaching platform based on large models. In *Proceedings of the 2024 International Symposium on Artificial Intelligence for Education* (pp. 409–415). https://doi.org/10.1145/3700297.3700367

Zhang, D., Zhao, X., & Zhou, G. (2024). Application of generative AI in the teaching of human resource management course for medical and health management majors. In *2024 14th international conference on Information Technology in Medicine and Education (ITME)* (pp. 735–739). IEEE. https://doi.org/10.1109/ITME63426.2024.00149