Advances in Medical Education and Practice

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Hotspot Evolution and Future Prospects of Large Language Models in Medical Education: A Bibliometric Analysis

Can Huang, Wei Liu

Department of Pharmacy, Beijing You'an Hospital Affiliated to Capital Medical University, Beijing, 100069, People's Republic of China

Correspondence: Wei Liu, Email 252691159@qq.com

Background: Advances in AI and NLP have popularized large language models (LLMs) in medical education. Post-2022 research has proliferated yet remains fragmented; no comprehensive bibliometric mapping systematically outlines this field's global layout, collaboration networks and thematic evolution.

Objective: To clarify publication trends, core contributors, collaboration patterns, research hotspots and evolutionary frontiers of LLMs in medical education via bibliometric analysis, and deliver targeted insights for educational practice.

Methods: This is a systematic bibliometric study. We collected 2016–March 2026 peer-reviewed English papers from WoSCC and Scopus. Metadata were standardized and analyzed with the Bibliometrix R package and VOSviewer to generate publication statistics, collaboration networks, citation metrics and keyword cluster maps.

Results: In total, 1991 papers were included, with an annual growth rate of 59.04%. The US (28.5%) and China (15.9%) led global outputs; Harvard, the University of Toronto and top Chinese scholars dominated contributions, and JMIR Medical Education served as the core journal. Five thematic clusters were identified: education ethics, LLM performance, patient education, clinical reasoning and intelligent assessment. Research priorities shifted from basic neural network exploration to privacy and standardized large-scale LLM deployment after 2024.

Conclusion: LLM medical education research grows rapidly but suffers insufficient empirical evidence, unbalanced themes and delayed governance frameworks. Based on our bibliometric findings, we propose practical optimization schemes: strengthen inter-disciplinary research, build matched risk supervision, define LLMs as teaching assistants, and prioritize cohort trials, specialized medical LLMs and unified ethical standards.

Academic Contributions: This study provides the latest full-spectrum quantitative mapping of the field, fills gaps in systematic literature review, and offers actionable references for medical educators, curriculum developers and policymakers to realize safe, high-quality LLM integration into medical education.

Keywords: large language models, artificial intelligence, medical education, bibliometric analysis, research trends

Introduction

The rapid breakthroughs in artificial intelligence (AI) and natural language processing (NLP) have catalyzed the widespread integration of large language models (LLMs) into nearly every aspect of medical education.[1–4] Before the 2022 LLM boom, early AI medical education research mainly focused on simple machine learning-assisted question banks and basic online tutoring tools, lacking intelligent conversational capabilities and systematic clinical scenario simulation capacity, forming a clear baseline gap compared with current generative AI applications. Since the launch of ChatGPT in late 2022, the field has undergone explosive growth, with LLMs widely adopted for clinical reasoning training, virtual patient simulation and automated teaching assessment.[5–8]

Nevertheless, LLM adoption brings prominent risks including algorithmic bias, factual hallucinations, student privacy leakage and lagging ethical governance mechanisms, which have not been systematically sorted out in existing scattered

https://doi.org/10.2147/AMEP.S625578
Received: 18 May 2026
Accepted: 1 July 2026
Published: 9 July 2026

studies.[9–11] Prior literature exploring LLM embedding into medical curricula mostly adopts narrow narrative reviews, focusing only on single teaching links or partial ethical disputes; few studies systematically comb cross-national collaboration, thematic evolution and field structural defects, resulting in highly fragmented research outputs without a holistic field roadmap for educators and policymakers.

Bibliometric analysis serves as a quantitative mapping tool with replicability and macroscopic visualization advantages, which can efficiently sort massive literature's intellectual structure and collaborative networks; yet it also carries limitations such as dependence on indexed metadata and unavoidable subjectivity in keyword clustering. This method is especially appropriate for this fast-expanding LLM medical education field, as traditional narrative reviews cannot achieve comprehensive panoramic sorting.[12,13]

Against the above context, this study's core purpose is to conduct a complete bibliometric mapping of global LLM medical education literature from 2016 to March 2026. Its positive academic contributions are threefold: first, it supplements the missing macroscopic overview of this emerging field; second, it identifies unbalanced thematic development and insufficient high-quality empirical evidence; third, it provides targeted optimization directions for safe, standardized LLM curriculum integration. The specific research objectives are as follows: Describe the temporal publication trends; Identify major contributing countries, institutions and authors; Visualize international and institutional collaboration networks; Analyze keyword clusters to reveal research hotspots; Explore evolutionary trends and propose future directions.

Methods

Search Strategy

This study adopts a systematic bibliometric descriptive research design, and the complete research workflow includes database retrieval, literature screening, metadata extraction, keyword unification, and bibliometric statistical visualization analysis.

In March 2026, we retrieved literature from two core databases: Web of Science Core Collection (WoSCC) and Scopus. WoSCC boasts complete standardized metadata and is widely recognized as the mainstream database for bibliometric research.[14,15] The retrieval formula combined terms of large language models and medical education:

(Topic: ("Large language models" OR "LLMs" OR "Big language models" OR "Artificial intelligence language models" OR "AI language models")) AND (Topic: ("Medical Education" OR "Medical Teaching" OR "Undergraduate Medical Education" OR "Graduate Medical Education" OR "Continuing Medical Education")).

The identical retrieval logic was applied to Scopus. The retrieval time range was set as 2016 to March 2026, covering the embryonic stage of AI medical education to the post-ChatGPT explosive research phase, with rational temporal coverage. Inclusion criteria: English original articles and reviews only; conference papers, preprints and grey literature were excluded. The full screening procedure is shown in Figure 1.

Data Extraction

The full records and cited references of the included publications were exported in tab-delimited text format. The bibliometric parameters that were extracted covered various aspects, such as titles, abstracts, keywords, authors, affiliations, countries or regions, publication year, journal names, and references.

To guarantee the accuracy of all the data, a double-verification process was carried out. In case there were any inconsistencies found during this process, they were resolved by re-examining the original articles.

When it came to the consolidation of keywords, multiple criteria were utilized. These included semantic equivalence, which meant that terms with the same meaning were grouped together; morphological variations, taking into account different forms of the same word; abbreviations and their corresponding full forms to unify the different ways of expressing the same term; domain-specific terminological standards to follow the proper usage within the relevant field; and contextual links where certain terms frequently appeared together and were closely related in context. By adopting these criteria, the aim was to merge the terms that represented the same core concept. This, in turn, enhanced the replicability and clarity of the subsequent analysis.

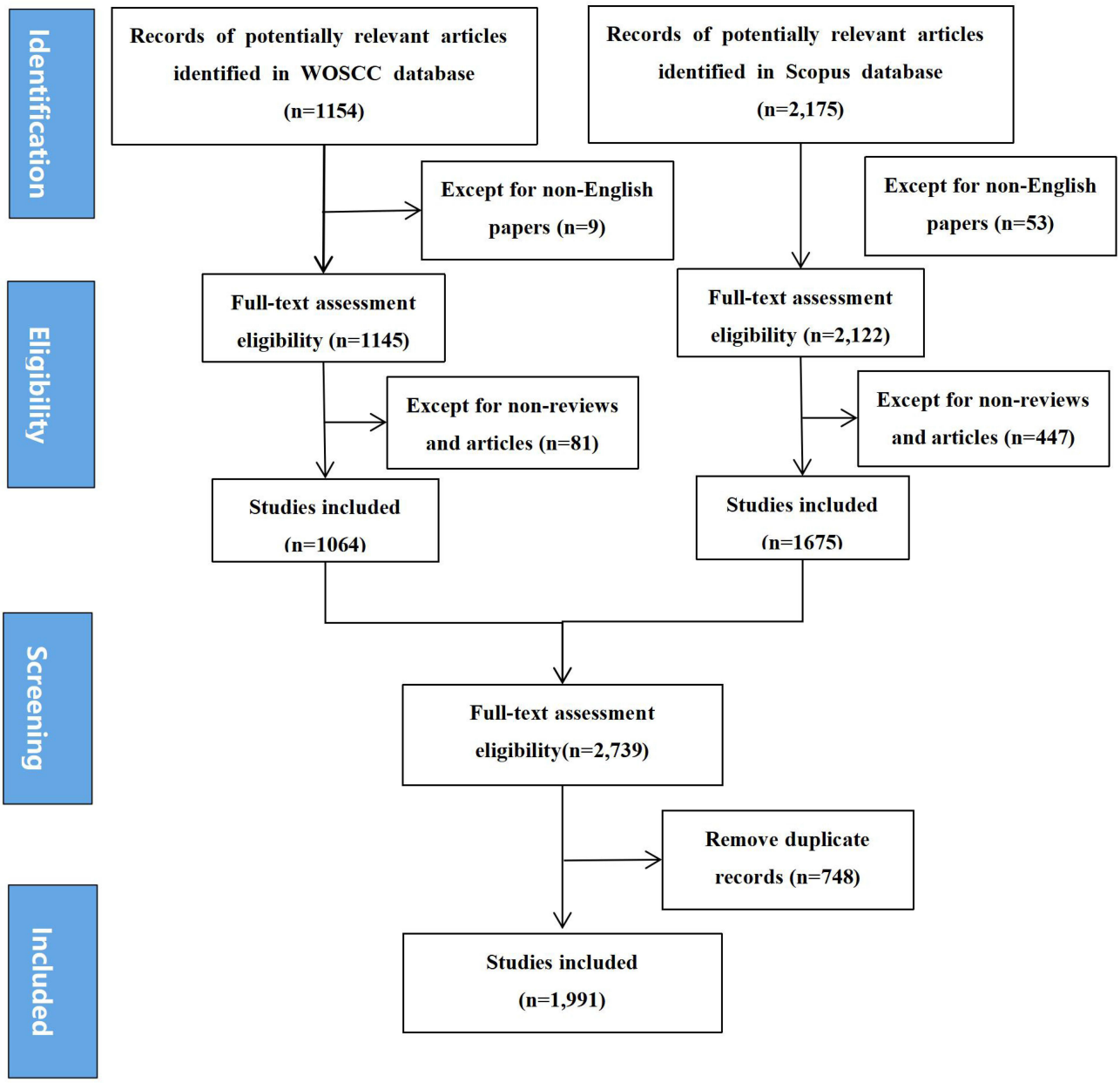


Figure 1 Flow diagram for the screening procedure. (Time range from 2016 to March 2026).

Data Analysis

In this study, the Bibliometrix package in R software (version 4.4.3) and VOSviewer (version 1.6.20) were employed for bibliometric analysis and the creation of scientific knowledge maps.[16,17]

The Bibliometrix R package played a significant role in analyzing several aspects. It was mainly utilized to examine the annual production of relevant literature, the production volume by different countries, the contributions made by authors over time, the local impacts of sources based on the H index, as well as the trending topics.[18,19]

VOSviewer is a powerful bibliometric tool. It was used to generate knowledge maps based on web data and enabled the visualization and exploration of these maps.[20] Specifically, in this study, due to its outstanding performance in clustering tasks, which is intuitive and clear, VOSviewer was leveraged for conducting clustering analyses of countries, institutions, journals, authors, citations, and keywords.[21]

https://doi.org/10.2147/AMEP.S625578

Both of these tools focus on analyzing co-occurrence. For instance, they analyze the instances where certain keywords appear together in publications. Through this analysis, they aim to map out the relationships among various elements and identify the natural clusters composed of tightly linked items. VOSviewer primarily visualizes these networks and clusters directly from the co-occurrence data by using distance and color. Meanwhile, the Bibliometrix package calculates the underlying matrices, provides statistical clustering methods, and detects trends through features like thematic evolution diagrams over time.

Results

Bibliometric Landscape Overview

A total of 1991 documents were retrieved from WoSCC and Scopus (Figure 1). Most of these publications were published after 2022. The retrieved literature spanned 912 distinct journals. Notably, the average annual growth rate of publications was 59.04%, and each document received an average of 20.86 citations. Additionally, the total number of contributing authors was 9554 (Table 1).

Publication Dynamics

Figure 2 illustrates the annual scientific production (number of articles) in the field of LLMs in medical education from 2016 to 2026 (with 2026 data reflecting only publications up to March of that year). From 2016 to 2021, the annual output remained at an extremely low and stable level, with only minor fluctuations and negligible growth. This period reflects the early exploratory phase of the field, characterized by limited research interest and foundational technological development. Beginning in 2022, the field entered a phase of exponential growth: the number of articles rose rapidly, accelerating sharply in 2023 and 2024, before reaching a clear peak in 2025, with the total number of published works approaching 1000.

Geographical and Institutional Output

National Productivity and Collaboration: Analysis of national publication output revealed contributions from 79 countries/regions. The country-specific distribution of publications is summarized in Table 2 and visualized in Figure 3. The United States was the most prolific contributor (n=568, 28.5% of total publications), followed by China (n=316, 15.9%), Germany (n=104, 5.2%), Canada (n=78, 3.9%), and the United Kingdom (n=65, 3.3%). To explore international research collaboration, the country co-authorship network was visualized using the country co-authorship

Table 1 Synopsis of Literature Search Outcomes

Account	Results
Main Information About Data	
Timespan	2016:2026
Sources (Journals)	912
Documents	1991
Annual Growth Rate %	59.04
Document Average Age	1.6
Average citations per doc	20.86
Document Contents	
Keywords Plus	5053
Author's Keywords	3799
Authors	
Authors	9554
Authors of single-authored docs	90
Document Types	
Article	1664
Review	327

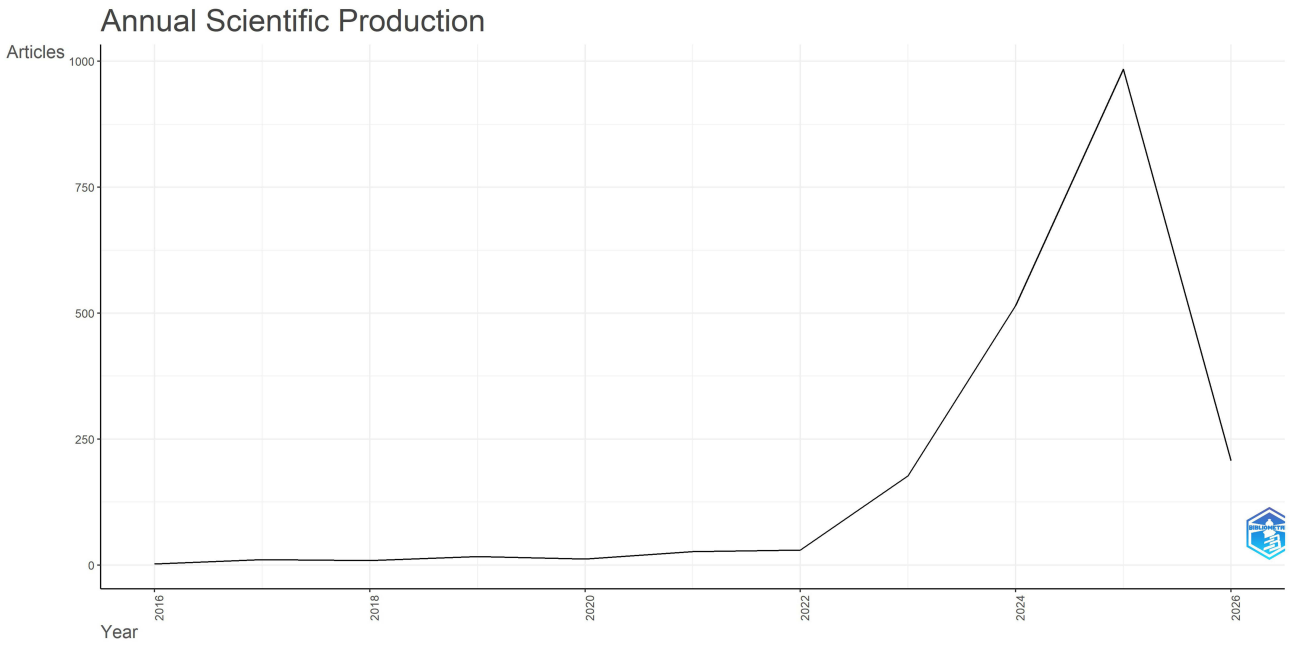


Figure 2 Distribution of yearly article outputs from 2016 to 2026 (with 2026 data reflecting only publications up to March of that year).

module in VOSviewer (Figure 4). A total of 45 productive countries/regions formed an interconnected collaborative network, among which the United States, the United Kingdom, China, Germany, and Canada emerged as prominent nodes with relatively thick connection lines. The United States exhibited the highest total link strength (TLS=244) and collaborated with 38 productive countries/regions.

Institutional Leadership: Among the top 10 most productive institutions (Table 3), Harvard University (USA) ranked first with 106 publications, closely followed by the University of Toronto (Canada) (n=97) and the University of California System (USA) (n=72). Additionally, Figure 5 depicts the collaboration network among the top 60 institutions. The Harvard Medical School node is the largest and most central in the network, indicating that it is the most prolific institution and a major hub for international collaboration.

Scholarly Ecosystem: Periodicals and Researchers

Journal Influence: A total of 1991 publications from 912 journals were included in this study. The top 10 journals in terms of publication volume, along with their 2024 Impact Factors (IF), are summarized in Table 4.[16] Notable journals among these include JMIR Medical Education (n=94, IF=3.2), Journal of Medical Internet Research (n=65, IF=6), and BMC Medical

Table 2 The 10 Most Productive Nations in Research on LLMs in Medical Education

Rank	Country	Articles	Articles %	SCP	MCP	MCP %
1	USA	568	28.5	515	53	9.3
2	China	316	15.9	270	46	14.6
3	Germany	104	5.2	84	20	19.2
4	Canada	78	3.9	58	20	25.6
5	United Kingdom	65	3.3	54	11	16.9
6	India	58	2.9	52	6	10.3
7	Italy	52	2.6	43	9	17.3
8	Japan	49	2.5	41	8	16.3
9	Australia	46	2.3	39	7	15.2
10	Korea	41	2.1	36	5	12.2

https://doi.org/10.2147/AMEP.S625578

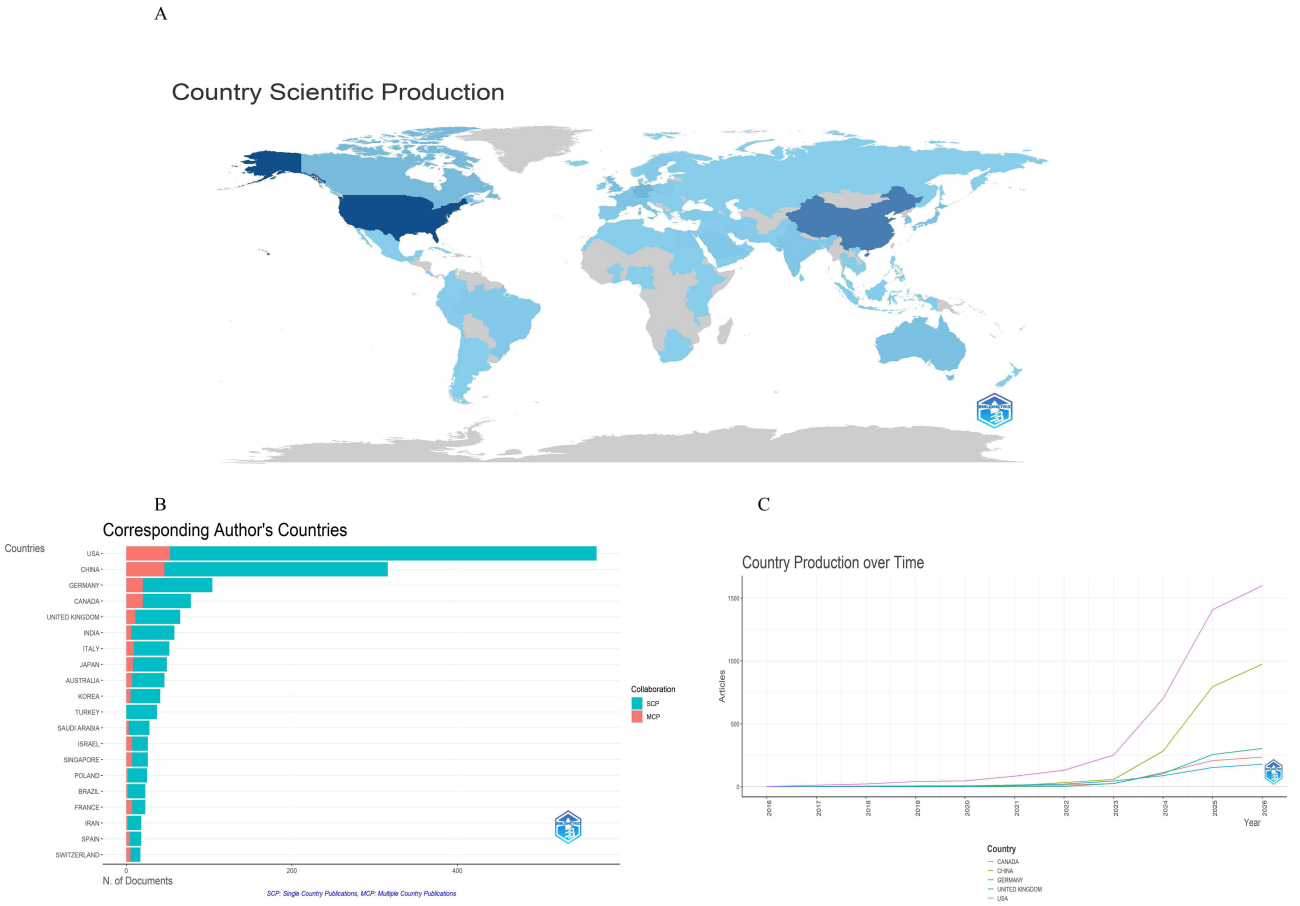
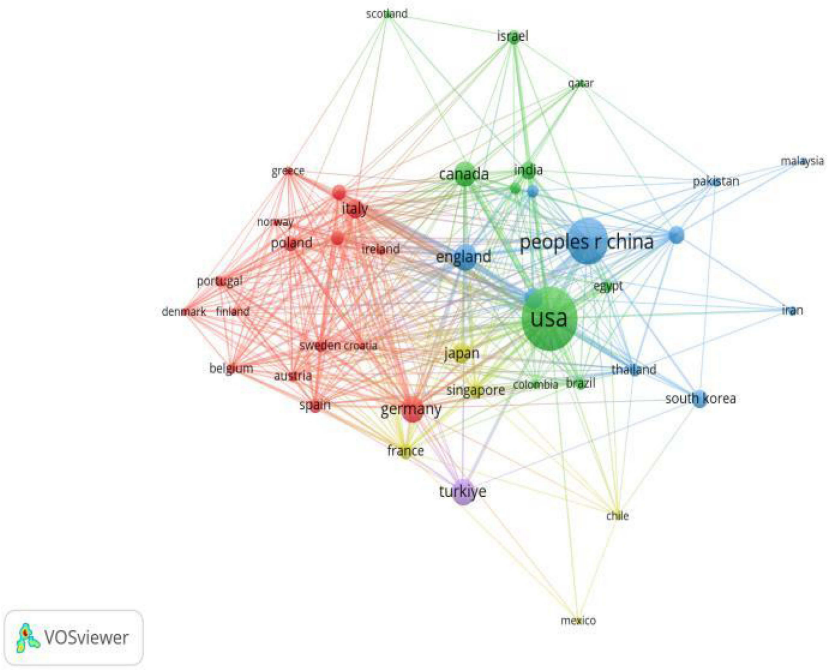


Figure 3 National Contributions to Research on LLMs in Medical Education. (**A**) Geographical map depicting national scientific output: darker blue shades indicate higher publication volumes; (**B**) Article publication volumes across countries from 2016 to 2026; (**C**) Countries of corresponding authors. SCP: Single Country Publications; MCP: Multiple Country Publications.

Education (n=55, IF=3.2). To evaluate journal influence, the Bibliometrix R package was employed, with measurements based on the H-index. JMIR Medical Education exhibited the highest H-index (Figure 6A). Temporal publication trends for these top journals are depicted in Figure 6B. For co-citation analysis, VOSviewer was used to analyze source journals (defined as those with ≥150 citations). Thirty journals were identified based on total link strength (Figure 6C), with the top three being JMIR Medical Education, Journal of Medical Internet Research and NPJ Digital Medicine.

Key Authors and Collaborative Networks: A total of 9554 authors worldwide have contributed to publications in this field. For this analysis, the top 10 authors by publication output were identified as key authors. Table 5 presents detailed metrics-including h-index, g-index, and m-index-calculated over the 10-year study period (2016–2026). Of the 10 authors listed in Table 5, nine are Chinese researchers, reflecting the strong contribution of Chinese scholars to this field. Among them, Zhang Y was the most productive with 35 cumulative publications, Wang Y exhibits the highest citation impact with a total of 2450 citations, and Li J and Zhang Y share the highest h-index of 14, indicating the strongest combined research productivity and long-term academic influence. Collectively, these three authors are identified as the key contributors in the dataset, demonstrating outstanding performance in research output, citation influence, and sustained academic impact. Figure 7 illustrates the temporal evolution of research productivity and citation impact for the top contributing authors in the field. All authors exhibited a clear upward trend in annual output, with a notable acceleration in 2023–2024.

A



B

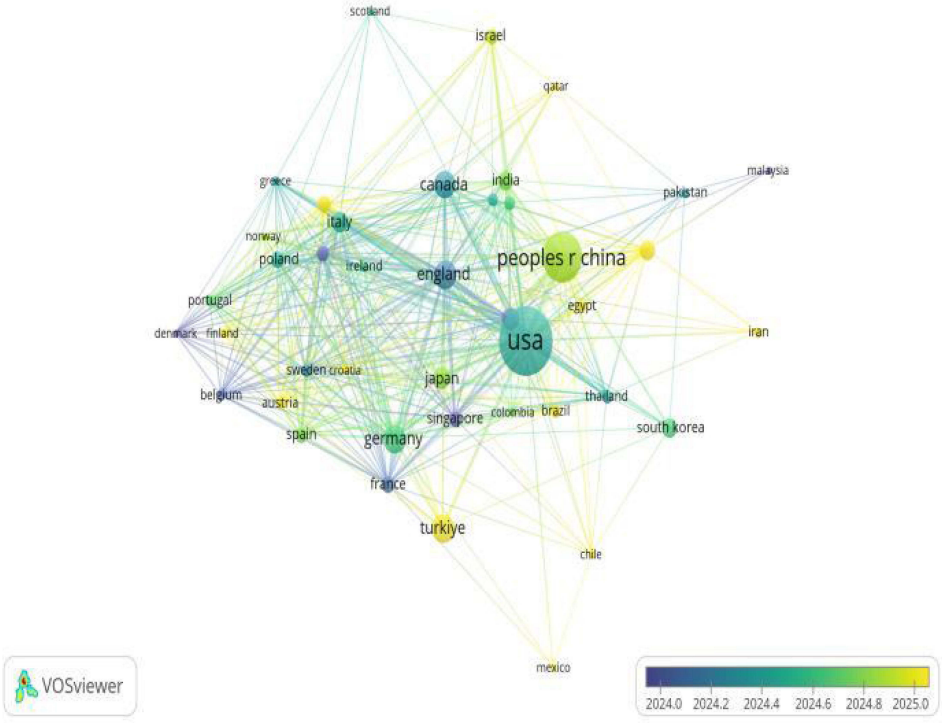


Figure 4 Analysis of Countries Involved in Research on LLMs in Medical Education. (**A**) Co-occurrence network visualization of countries in research on LLMs in medical education. Countries are clustered into five color-coded groups, with larger nodes indicating higher-productivity nations; (**B**) Countries colored by their average publication year, where blue represents earlier-stage contributors and yellow represents later-stage contributors.

https://doi.org/10.2147/AMEP.S625578

Table 3 The Top 10 Institutions with the Highest Productivity

Rank	Title of the Institution	Literature	Nation
1	Harvard University	106	USA
2	University of Toronto	97	Canada
3	University of California System	72	USA
4	Harvard Medical School	71	USA
5	Mayo Clinic	70	USA
6	Harvard University Medical Affiliates	64	USA
7	Sichuan University	61	China
8	University System OF Ohio	61	USA
9	Stanford University	53	USA
10	University OF California	48	USA

Keywords and Research Frontiers

Keyword Standardization

A thesaurus was employed (Supplementary Table 1) to enable more precise counting of keyword occurrence frequencies.

Clusters of Keyword Co-Occurrence

Figure 8A visualized 50 high-frequency keywords meeting a co-occurrence threshold of ≥15. The top 10 keywords by co-occurrence frequency are: artificial intelligence (n=665); large language models (n=535); ChatGPT (n=407); medical education (n=356); natural language processing (n=105); patient education (n=100); chatbots (n=76); machine learning (n=64); performance (n=47); and medical students (n=43). Five major thematic clusters were identified via VOSviewer:

(1) Red Cluster: medical education subjects, clinical practice, talent training, and ethical oversight.

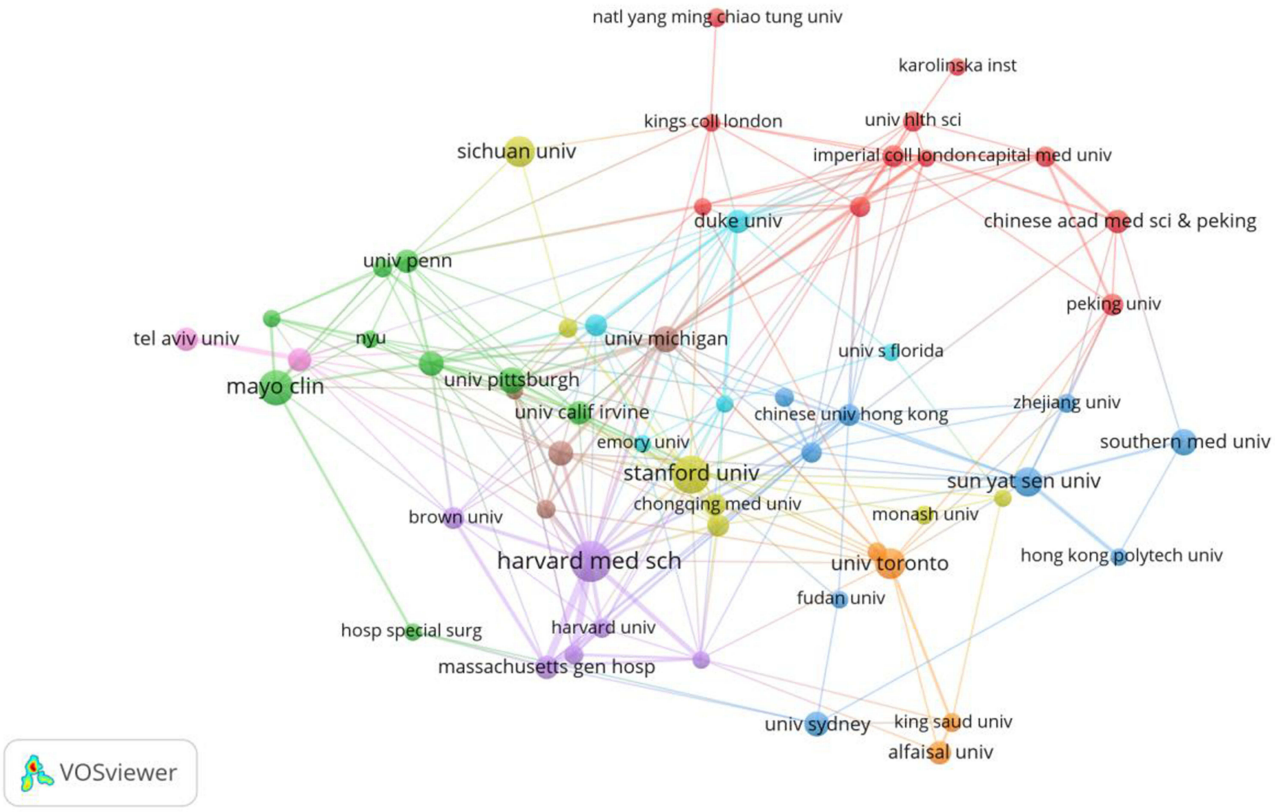


Figure 5 Clustering analysis of collaborative ties among institutions with over 5 publications.

Table 4 The Top 10 Journals with the Highest Publication Output

Rank	Periodical	Publication Counts	Citation Counts	Impact Factor	Quartile Ranking
1	JMIR Medical Education	94	4791	3.2	Q1
2	Journal of Medical Internet Research	65	1856	6	Q1
3	BMC Medical Education	55	708	3.2	Q1
4	Scientific Reports	33	349	3.9	Q1
5	Medical Teacher	28	320	4.4	Q1
6	Academic Medicine	21	394	5.2	Q1
7	International Journal of Medical Informatics	16	387	4.1	Q1
8	Journal of Medical Systems	14	378	5.7	Q1
9	British Journal of Ophthalmology	11	316	3.5	Q1
10	Journal of The American Medical Informatics Association	9	310	4.6	Q1

(2) Green Cluster: LLMs themselves, technical performance, digital health, and quality evaluation.

(3) Yellow Cluster: patient health education, health information dissemination, and readability optimization.

(4) Blue Cluster: advanced clinical reasoning, cutting-edge model iteration, and intelligent decision support.

(5) Purple Cluster: medical assessment, natural language processing, and standardized education evaluation.

Emerging Research Frontiers

Regarding research trends, emerging key themes in this field (Figure 8B) include recognition, privacy, scale, responses, burden, ChatGPT, artificial intelligence, education, and performance. The development of the field can be divided into four stages:

2018–2020: Focused on neural networks, basic language models, and resident training.

2020–2022: Expanded to technology application, therapy, and knowledge construction.

2022–2024: Dominated by ChatGPT, artificial intelligence, educational impact, and cancer teaching.

2024–2026: Shifted toward privacy protection, large-scale application, information quality, and recognition technology.

Discussion

Research Growth Trajectory and Driving Forces

The research on LLMs in medical education has shown an exponential upward trend, with an annual growth rate as high as 59.04% from 2016 to 2026, far exceeding most subfields of medical education. This explosive expansion is not merely quantitative but reflects a profound paradigm shift driven by the advent of generative AI, particularly ChatGPT, which has catalyzed a transition from theoretical exploration to real-world educational implementation.[22,23]

From 2018 to 2020, the field was in the technological nascent stage, focusing on traditional neural networks, basic language models, and resident training without mature generative models.[24] From 2020 to 2022, research expanded to technical applications, clinical therapy, and knowledge construction, laying a foundation for subsequent integration. Since the release of ChatGPT at the end of 2022, the field entered an explosive growth stage from 2022 to 2024, with research hotspots rapidly concentrating on model performance, educational impact, and clinical teaching.[25,26] After 2024, the focus shifted to in-depth regulation and addressing frontier challenges, emphasizing privacy protection, large-scale implementation, information quality, and readability. This trajectory mirrors the global development of generative AI, underscoring the tight coupling between technical breakthroughs and educational research priorities.[27,28] Notably, the inflection point in late 2022 marks the moment when LLMs evolved from niche computational tools into mainstream educational instruments, validating the transformative potential of AI in reshaping medical pedagogy.[29]

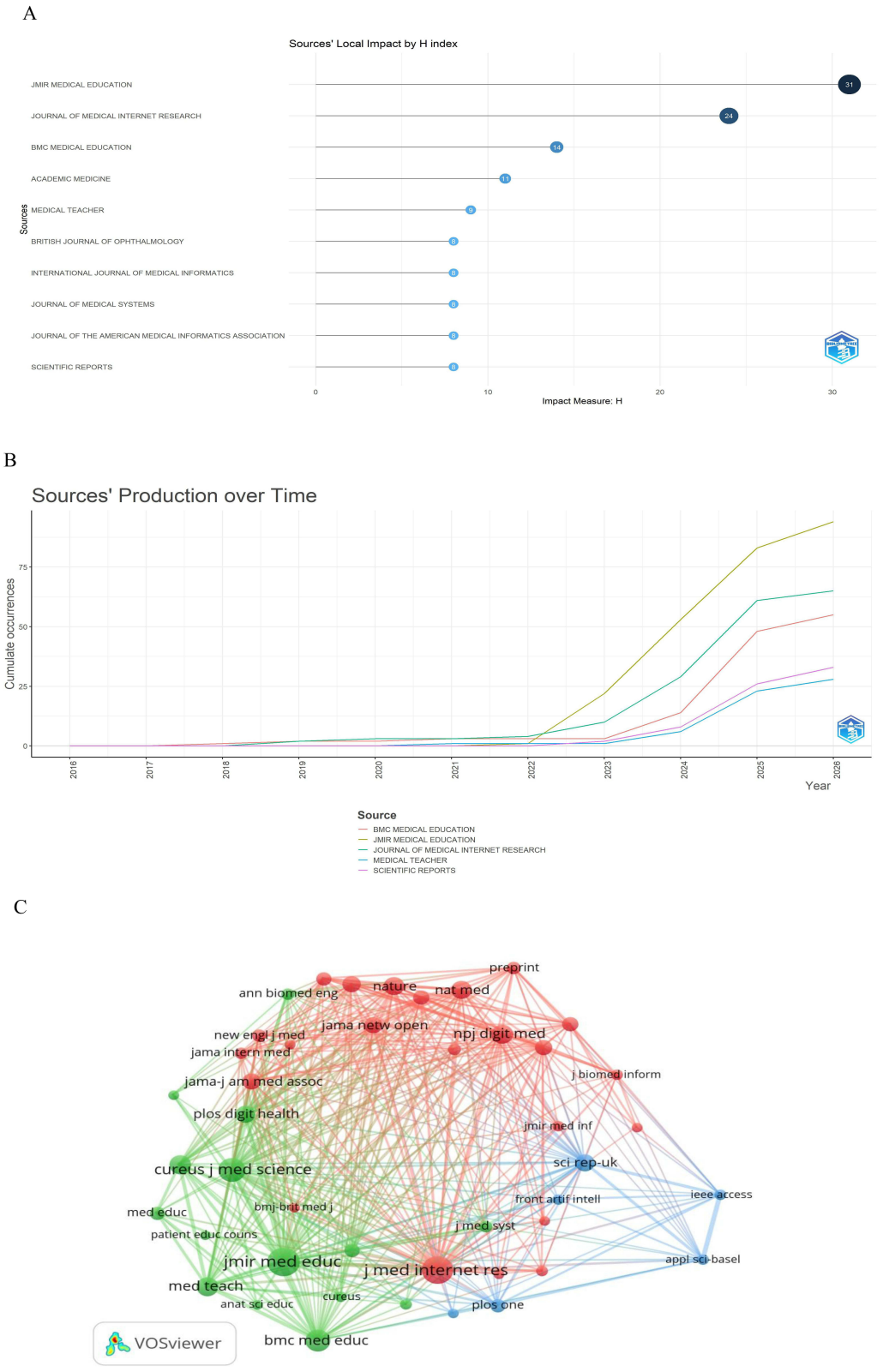


Figure 6 Overview of Leading Journals and Co-Citation Analysis. (**A**) Timeline distribution of publications in the top 10 high-output journals; (**B**) Annual article volumes in journals spanning 2016–2026; (**C**) Co-cited journals in research on LLMs in medical education.

Table 5 Publication and Citation Metrics for Key Authors

Author	h_Index	g_Index	m_Index	Total Citation Count	Cumulative Publications
Li J	14	29	2	892	29
Zhang Y	14	35	2.8	2092	35
Chen Y	10	19	2.5	406	33
Kim J	10	20	2.5	410	20
Wang Y	10	32	2	2450	32
Li Y	9	15	3	246	26
Wang X	9	27	1.286	2381	27
Wu Y	9	17	1.125	1959	17
Chen J	8	14	0.889	207	24
Liu J	8	14	2	203	17

Geographic and Institutional Contributions

Geographically, the global research landscape is dominated by the United States and China, which together account for 44.4% of total publications. The United States leads with 568 articles (28.5%), demonstrating robust single-country publications (SCP=515) and extensive international collaboration (multi-country publications, MCP=53; total link strength, TLS=244), positioning it as a central hub in the global collaborative network. This leadership aligns with longstanding U.S. investments in medical informatics and interdisciplinary AI research, as evidenced by institutions like Harvard University and the University of California System-both ranking among the top three most productive global entities. China ranks second with 316 publications (15.9%), exhibiting rapid growth in recent years; however, its MCP ratio (14.6%) is relatively modest compared to Western nations, indicating untapped potential for deeper cross-border partnerships. This dual-center structure reflects both America's historical dominance in medical education technology and China's strategic push into AI-enabled "New Medical Sciences" under national policy frameworks.[30]

Institutional analysis further reveals that North American academic medical centers dominate productivity, with Harvard University, the University of Toronto, and the University of California System leading the global rankings. Notably, Sichuan University represents the most influential Chinese institution, signaling the emergence of Asian centers in this globally competitive field. The intensive collaboration among these top-tier universities highlights the importance of resource concentration and interdisciplinary synergy-for example, joint projects between engineering and medical faculties often yield the most impactful LLM applications in simulation-based training or clinical reasoning tools.

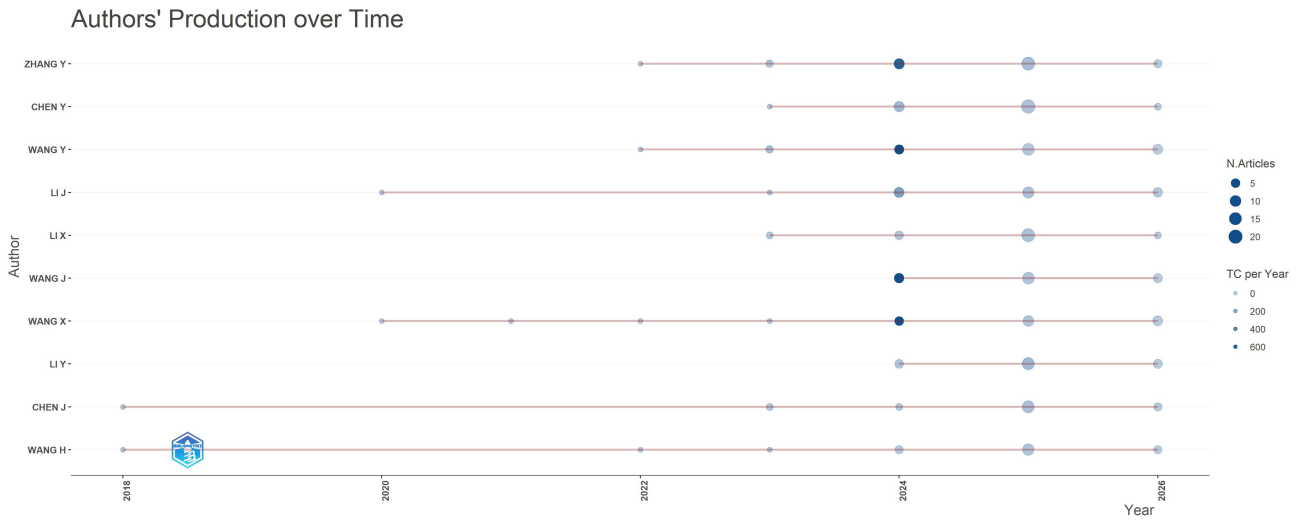


Figure 7 Depiction of Authors' Productivity and Citation Impact Over Time. Larger nodes signify a greater number of publications per year, while darker nodes indicate higher annual citation impact.

https://doi.org/10.2147/AMEP.S625578

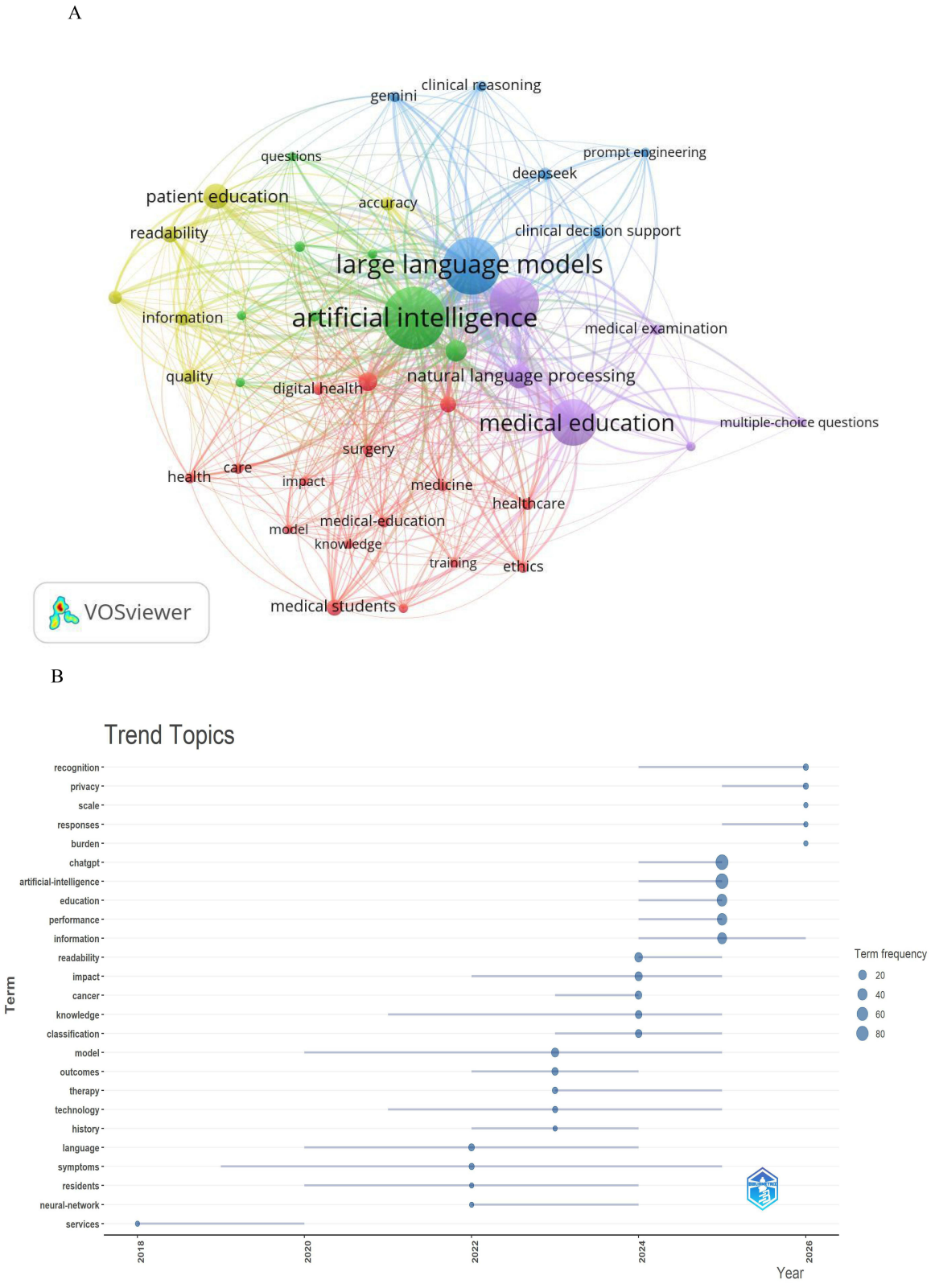


Figure 8 Keyword Analysis of LLM-Related Medical Education Research. (**A**) Keyword co-occurrence network visualization for studies on LLMs in medical education. Keywords are grouped into five clusters by color, with larger nodes indicating more frequently occurring terms; (**B**) Keyword trends and their evolution in LLM-related medical education research.

However, a critical gap remains in partnerships between research institutions and frontline clinical teaching hospitals: most studies originate from university labs, with limited input from clinicians who oversee day-to-day medical training, risking a disconnect between technical innovation and real-world educational needs.[31]

Leading Journals and Influential Authors

Scholarly communication in this domain is concentrated in high-impact journals specializing in medical education and informatics. JMIR Medical Education emerges as the core venue with the highest H-index, followed by the Journal of Medical Internet Research and BMC Medical Education-all Q1 journals with stable citation influence. This concentration confirms that LLM-based medical education has gained broad recognition within the international academic community, with journals proactively issuing dedicated calls for papers on AI integration since 2023. Author-level metrics identify Li J, Zhang Y, and Wang Y as key contributors, excelling in publication volume, citation impact, and h-index. Remarkably, 9 out of the top 10 authors are Chinese, reflecting the substantial and growing contribution of Chinese scholars to this global field. Their active publishing period aligns with the post-2022 AI boom, illustrating the rapid formation of a high-impact, youthful researcher cohort driving disciplinary advancement.

This trend is further reinforced by the rise of Chinese-developed LLMs[32,33] which have demonstrated superior performance in Chinese medical licensing examinations and clinical dialogue tasks compared to general-purpose models like GPT-4. For example, Qwen-2.5 achieved the highest accuracy (89.8%) among tested models on the Chinese National Nursing Licensing Examination,[34] a result attributed to its training on a domain-specific Chinese medical dataset.

Thematic Clusters and Structural Deficiencies

Based on keyword co-occurrence analysis, five interrelated and clearly differentiated thematic clusters were summarized in this study. These clusters jointly constitute the complete research framework of LLM application in medical education, covering teaching practice, technical verification, public education, clinical training, and academic evaluation. Each cluster has independent research focuses and practical values, while sharing inherent limitations that restrict the high-quality development of this field.

Red Cluster: Medical Education Ontology, Talent Training, and Ethical Governance

As the largest research cluster in this field, this module takes medical students and clinical practice teaching as the core research objects. It mainly explores the full-cycle talent training mode of medical majors and the ethical risk management in the application of AI technology. This cluster closely fits the essential attributes of medical education, attaches importance to clinical practical orientation, and puts forward early warnings on ethical risks such as algorithmic bias and data privacy.[35,36] Nevertheless, this research branch has obvious deficiencies. Most ethical discussions remain at a macroscopic theoretical level, lacking hierarchical management systems and operable review standards suitable for medical teaching scenarios.[37,38] In addition, existing research is dominated by short-term cross-sectional surveys, with a scarcity of long-term longitudinal tracking data.[39] Moreover, there is a lack of differentiated training schemes tailored to students at different learning stages and various medical specialties.[40]

Green Cluster: Model Performance, Digital Health, and Quality Control

With LLMs as the core, this cluster verifies model accuracy, information reliability, and digital health transformation. It provides the technical foundation for educational application.[41,42] However, overemphasis on accuracy ignores critical risks such as hallucination, logical loopholes, and lack of interpretability;[43,44] evaluation indicators are too generalized to adapt to high-risk medical scenarios; most performance tests are conducted in ideal laboratory environments, lacking verification in real complex teaching scenarios.[45–47]

Yellow Cluster: Patient Health Education, Health Literacy, and Readability Optimization

This cluster extends medical education from campus training to public health education, focusing on patient education, popular science communication, and readability improvement.[48–50] It expands the boundary of medical education and promotes health equity.[51] Yet risk prevention of misleading information is seriously insufficient; excessive pursuit of

efficiency ignores emotional communication and humanistic care; digital equity and accessibility for vulnerable groups are rarely discussed.[52,53]

Blue Cluster: Advanced Clinical Reasoning, Frontier Models, and Prompt Engineering

As the most cutting-edge and fastest-growing cluster, it targets clinical reasoning, decision support, new models (Gemini, DeepSeek), and prompt engineering. It leads the transformation from knowledge assistance to high-level ability training. Nevertheless, some studies overstate model efficacy and blur the boundary between assistance and independent decision-making; prompt engineering has not been incorporated into the formal curriculum; and there is a lack of unified and authoritative evaluation frameworks for multi-model comparison.[54,55]

Purple Cluster: Medical Assessment, NLP, and Standardized Evaluation

Focusing on the terminal link of education, this cluster explores intelligent assessment, automated scoring, and objective question evaluation. It greatly reduces teachers' workload. However, the evaluation dimension is seriously imbalanced, lacking assessment of humanistic care, practical skills, and doctor-patient communication;[56,57] it brings academic integrity risks such as cheating.[58]

Structural Deficiencies in the Current Field

Unbalanced Thematic Development

At present, research hotspots are excessively concentrated on technical exploration and clinical application, while research on humanistic literacy cultivation, long-term educational impact, and public health equity is insufficient.[59–61] Each thematic cluster develops independently, and interdisciplinary and cross-cluster collaborative research is scarce, failing to form a systematic and integrated research system.[2,62]

Technology Advances Faster Than Governance

The updating speed of LLM application technology is far ahead of the construction of supporting governance systems.[10,63] Research on ethical norms, data privacy protection, and long-term potential risks lags behind practical application by approximately two years, forming an unhealthy development pattern of "technology iteration first and governance supervision later".

Insufficient High-Quality Empirical Research

Existing literature is dominated by descriptive analysis and subjective opinion papers. High-level evidence-based research such as randomized controlled trials and long-term cohort tracking studies is seriously inadequate. Most conclusions lack rigorous empirical verification, which reduces the reliability and practical guiding value of relevant research results.[64,65]

More Criticism and Less Construction

Numerous studies have repeatedly pointed out the ethical risks, technical defects, and application limitations of LLMs, but few scholars propose targeted, operable optimization schemes.[66,67] There is a serious lack of standardized institutional frameworks, personalized curriculum plans, and complete supervision norms for LLM application in medical education.[68,69]

Risk of Deviating from Medical Education Orientation

Some studies overly focus on technical indicators such as model accuracy and response speed, ignoring the core educational goal of medical talent training. The patient-centered service concept and clinical competency-oriented training orientation are weakened, leading to the deviation of partial technical research from the essential purpose of medical education.[70,71]

Improvement Strategies and Future Research Directions

In view of the above structural deficiencies, this study puts forward targeted improvement strategies from the perspective of balanced and sustainable development, and clarifies high-value research directions for subsequent studies.

Balance the Thematic Layout

Researchers should attach importance to weak research directions including ethical governance, long-term educational effects, and digital health equity.[72,73] It is necessary to break the independent development barrier of each cluster and carry out interdisciplinary integrated research.[74]

Build a Synchronized Governance System

Prioritize risk prediction and institutional system construction to keep the governance pace consistent with technological updates. Strictly incorporate data privacy protection, algorithmic bias detection, and information security assessment into the technology access threshold.[75] A hierarchical risk management mechanism should be established to realize classified supervision of LLM large-scale application in different medical teaching scenarios.[76]

The Positioning of LLMs in Medical Education

Clarify the auxiliary positioning of LLMs in medical education rather than a substitute for manual teaching.[77] Develop virtual clinical cases and intelligent simulation training systems to improve students' practical operation ability.[78]

Future Research

To further promote the standardized development of this field, subsequent research can focus on four key directions: horizontal performance comparison of various advanced medical large models, long-term cohort research on the educational impact of LLMs, customized development of professional medical LLMs, and construction of unified standardized AI medical ethics system.

Limitations

Several limitations warrant consideration when interpreting the findings of this study:

1. Database Bias: This analysis only includes publications indexed in the WoSCC and Scopus, potentially underrepresenting non-English literature and regional journals.
2. Subjectivity in Clustering: Despite rigorous keyword standardization, semantic clustering and thematic classification retain a degree of subjectivity, which may affect the granularity of topic mapping.
3. Exclusion of Gray Literature: Conference abstracts, dissertations, preprints, and gray literature were excluded from the analysis, possibly omitting early-stage innovations and practical implementations in clinical teaching settings.
4. Bibliometric Metrics Limitation: Bibliometric metrics alone cannot capture the pedagogical quality, clinical utility, or long-term educational impact of included studies.

Notwithstanding these constraints, the large sample size (1991 publications), systematic screening protocol, and comprehensive visualization methods enhance the reliability and representativeness of our findings.

Conclusion

LLM medical education research is growing exponentially at an annual rate of 59.04%, with the United States and China accounting for 44.4% of global publications and the field structured around five core thematic clusters. While it delivers transformative educational opportunities ranging from personalized learning tools to virtual patient simulations, it still faces prominent challenges including insufficient high-level empirical evidence, unbalanced thematic development and lagging governance frameworks. This study makes a key academic contribution by providing the latest dual-database field mapping up to March 2026 and quantifying previously underrecognized structural imbalances in the field. Based on these statistical findings, we put forward three actionable targeted suggestions: integrating prompt engineering into formal medical curricula, prioritizing randomized controlled trials evaluating LLM educational outcomes, and establishing a graded risk supervision system for educational AI applications. Future research should balance innovation and safety, and align technological progress with ethical governance to advance responsible, high-quality LLM integration into medical education.

Data Sharing Statement

No datasets were generated or analysed during the current study.

Funding

The authors stated that no funding was involved in the work presented in this article.

Disclosure

The authors declare no competing interests in this work.

References

1. O'Leary K. LLMs get a medical education. *Nat Med*. 2023;2023:d41591.
2. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Medical Education*. 2024;58(11):1276–1285. doi:10.1111/medu.15402
3. Iqbal U, Tanweer A, Rahmanti AR, et al. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. *J Biomed Sci*. 2025;32(1):45. doi:10.1186/s12929-025-01131-z
4. Rao AS, Kim J, Mu A, et al. Synthetic medical education in dermatology leveraging generative artificial intelligence. *Npj Digit Med*. 2025;8:247. doi:10.1038/s41746-025-01650-x
5. Zhui L, Fenghe L, Xuehu W, et al. Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education: viewpoint. *J Med Internet Res*. 2024;26:e60083. doi:10.2196/60083
6. Ong JCL, Chang SY-H, William W, et al. Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*. 2024;6(6):e428–e432. doi:10.1016/S2589-7500(24)00061-X
7. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *Npj Digit Med*. 2024;7(1):183. doi:10.1038/s41746-024-01157-x
8. Hassan W, Duarte AE. Bibliometric analysis: a few suggestions. *Current Problems in Cardiology*. 2024;49(8):102640. doi:10.1016/j.cpcardiol.2024.102640
9. Ninkov A, Frank JR, Maggio LA. Bibliometrics: methods for studying academic publishing. *Perspect Med Educ*. 2021;11(3):173–176. doi:10.1007/S40037-021-00695-4
10. De Granda-Orive JI, Alonso-Arroyo A, Roig-Vázquez F. ¿Qué base de datos debemos emplear para nuestros análisis bibliográficos? Web of Science versus SCOPUS. *Archivos de Bronconeumología*. 2011;47(4):213. doi:10.1016/j.arbres.2010.10.007
11. Falagas ME, Pitsouni EI, Malietzis GA, et al. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *THE FASEB Journal*. 2008;22(2):338–342. doi:10.1096/fj.07-9492LSF
12. Zhang L, Zheng H, Jiang S-T, et al. Worldwide research trends on tumor burden and immunotherapy: a bibliometric analysis. *International Journal of Surgery*. 2024;110(3):1699–1710. doi:10.1097/JS9.0000000000001022
13. Bukar UA, Sayeed MS, Razak SFA, et al. A method for analyzing text using VOSviewer. *MethodsX*. 2023;11:102339. doi:10.1016/j.mex.2023.102339
14. Zhang X-D, Zhang Y, Zhao Y-Z, et al. Autoimmune pancreatitis: a bibliometric analysis from 2002 to 2022. *Front Immunol*. 2023;14:1135096. doi:10.3389/fimmu.2023.1135096
15. Cheng K, Guo Q, Shen Z, et al. Frontiers of ferroptosis research: an analysis from the top 100 most influential articles in the field. *Front Oncol*. 2022;12:948389. doi:10.3389/fonc.2022.948389
16. Devos P, Ménard J. Trends in Worldwide Research in Hypertension Over the Period 1999–2018: a Bibliometric Study. *Hypertension*. 2020;76(5):1649–1655. doi:10.1161/HYPERTENSIONAHA.120.15711
17. Van Eck NJ, Waltman L. Citation-based clustering of publications using CitNetExplorer and VOSviewer. *Scientometrics*. 2017;111(2):1053–1070. doi:10.1007/s11192-017-2300-7
18. Chen SF, Alyakin A, Seas A, et al. LLM-assisted systematic review of large language models in clinical medicine. *Nat Med*. 2026;32(3):1152–1159. doi:10.1038/s41591-026-04229-5
19. Hersh W. Generative Artificial Intelligence: implications for Biomedical and Health Professions Education. *Annual Review of Biomedical Data Science*. 2025;8(1):355–380. doi:10.1146/annurev-biodatasci-103123-094756
20. Yang GR, Wang X-J. Artificial Neural Networks for Neuroscientists: a Primer. *Neuron*. 2020;107(6):1048–1070. doi:10.1016/j.neuron.2020.09.005
21. Zhang Y, Xie X, Xu Q. ChatGPT in Medical Education: bibliometric and Visual Analysis. *JMIR Med Educ*. 2025;11:e72356–e72356. doi:10.2196/72356
22. Hui Z, Zewu Z, Jiao H, et al. Application of ChatGPT-assisted problem-based learning teaching method in clinical medical education. *BMC Med Educ*. 2025;25(1):50. doi:10.1186/s12909-024-06321-1
23. Safranek CW, Sidamon-Eristoff AE, Gilson A, et al. The Role of Large Language Models in Medical Education: applications and Implications. *JMIR Med Educ*. 2023;9:e50945. doi:10.2196/50945
24. Zhui L, Yhap N, Liping L, et al. Impact of Large Language Models on Medical Education and Teaching Adaptations. *JMIR Med Inform*. 2024;12:e55933–e55933. doi:10.2196/55933
25. Benítez TM, Xu Y, Boudreau JD, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. *Journal of the American Medical Informatics Association*. 2024;31(3):776–783. doi:10.1093/jamia/ocad252
26. Zhao D, Zhang Y, Wang J, et al. Cardiovascular disease prevention in China: challenges and opportunities in the artificial intelligence-enabled digital health era. *Nat Rev Cardiol*. 2026;23:363–374. doi:10.1038/s41569-025-01222-2
27. Swanson K, Wu W, Bulaong NL, et al. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*. 2025;646:716–723. doi:10.1038/s41586-025-09442-9

28. Sandmann S, Hegselmann S, Fujarski M, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nat Med*. 2025;31:2546–2549. doi:10.1038/s41591-025-03727-2
29. Yang X, Li T, Su Q, et al. Application of large language models in disease diagnosis and treatment. *Chinese Medical Journal*. 2025;138:130–142. doi:10.1097/CM9.0000000000003456
30. Zhu S, Hu W, Yang Z, et al. Qwen-2.5 Outperforms Other Large Language Models in the Chinese National Nursing Licensing Examination: retrospective Cross-Sectional Comparative Study. *JMIR Med Inform*. 2025;13:e63731. doi:10.2196/63731
31. Wong EYT, Verlingue L, Aldea M, et al. ESMO guidance on the use of Large Language Models in Clinical Practice (ELCAP). *Annals of Oncology*. 2025;36:1447–1457. doi:10.1016/j.annonc.2025.09.001
32. Zohny H. Reframing 'dehumanisation': AI and the reality of clinical communication. *J Med Ethics*. 2025;jme–2025–111307. doi:10.1136/jme-2025-111307
33. Heinke A, Radgoudarzi N, Huang BB, et al. A review of ophthalmology education in the era of generative artificial intelligence. *Asia-Pacific Journal of Ophthalmology*. 2024;13:100089. doi:10.1016/j.apjo.2024.100089
34. Fareed M, Fatima M, Uddin J, et al. A systematic review of ethical considerations of large language models in healthcare and medicine. *Front Digit Health*. 2025;7:1653631. doi:10.3389/fdgth.2025.1653631
35. Zheng J, Shi C, Cai X, et al. Lifelong Learning of Large Language Model Based Agents: a Roadmap. *IEEE Trans Pattern Anal Mach Intell*. 2026;48(5):5552–5571. doi:10.1109/TPAMI.2025.3650546
36. Suresh S, Misra SM. Large Language Models in Pediatric Education: current Uses and Future Potential. *Pediatrics*. 2024;154(3):e2023064683. doi:10.1542/peds.2023-064683
37. Wu X, Huang Y, He Q. A large language model improves clinicians' diagnostic performance in complex critical illness cases. *Crit Care*. 2025;29 (1):230. doi:10.1186/s13054-025-05468-7
38. Mishra V, Lurie Y, Mark S. Accuracy of LLMs in medical education: evidence from a concordance test with medical teacher. *BMC Med Educ*. 2025;25(1):443. doi:10.1186/s12909-025-07009-w
39. Mehta S, Mehta N. Embracing the illusion of explanatory depth: a strategic framework for using iterative prompting for integrating large language models in healthcare education. *Medical Teacher*. 2025;47(2):208–211. doi:10.1080/0142159X.2024.2382863
40. Yu E, Chu X, Zhang W, et al. Large Language Models in Medicine: applications, Challenges, and Future Directions. *Int J Med Sci*. 2025;22 (11):2792–2801. doi:10.7150/ijms.111780
41. He Q, Tan Z, Niu W, et al. From algorithms to operating room: can large language models master China's attending anesthesiology exam? A cross-sectional evaluation. *International Journal of Surgery*. 2026;112(1):190–201. doi:10.1097/JS9.0000000000003406
42. Liu J, Chen T, Li S, et al. LLM-based pedagogical agent for ICU simulation instructor training: a quasi-experimental study. *Nurse Education Today*. 2026;157:106901. doi:10.1016/j.nedt.2025.106901
43. Daccache N, Zako J, Morisson L, et al. The applications of ChatGPT and other large language models in anesthesiology and critical care: a systematic review. *Can J Anesth/J Can Anesth*. 2025;72(6):904–922. doi:10.1007/s12630-025-02973-9
44. Rizkala T, Muench N, Hassan C, et al. Generative artificial intelligence for patient education material on gastric cancer prevention. *Endoscopy*. 2026;2026:2780.
45. Sousa-Pinto B, Vieira RJ, Marques-Cruz M, et al. Artificial Intelligence–Supported Development of Health Guideline Questions. *Ann Intern Med*. 2024;177(11):1518–1529. doi:10.7326/ANNALS-24-00363
46. Zaretsky J, Kim JM, Baskharoun S, et al. Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format. *JAMA Netw Open*. 2024;7(3):e240357. doi:10.1001/jamanetworkopen.2024.0357
47. Chen H, Zeng D, Qin Y, et al. Large language models and global health equity: a roadmap for equitable adoption in LMICs. *The Lancet Regional Health - Western Pacific*. 2025;63:101707. doi:10.1016/j.lanwpc.2025.101707
48. Lawson McLean A. Constructing knowledge: the role of AI in medical learning. *Journal of the American Medical Informatics Association*. 2024;31 (8):1797–1798. doi:10.1093/jamia/ocae124
49. Lu H, Lin Y, Li Z, et al. Toward fair medical advice: addressing and mitigating bias in large language model-based healthcare applications. *Artificial Intelligence in Medicine*. 2025;168:103216. doi:10.1016/j.artmed.2025.103216
50. Mansoor I, Abdullah M, Rizwan MD, et al. Reasoning with large language models in medicine: a systematic review of techniques, challenges and clinical integration. *Health Inf Sci Syst*. 2025;14(1):6. doi:10.1007/s13755-025-00403-0
51. Luo Z, Qiao Y, Xu X, et al. Cross sectional pilot study on clinical review generation using large language models. *Npj Digit Med*. 2025;8(1):170. doi:10.1038/s41746-025-01535-z
52. Idan D, Einav S. Primer on large language models: an educational overview for intensivists. *Crit Care*. 2025;29(1):238. doi:10.1186/s13054-025-05479-4
53. Simoni J, Urtubia-Fernandez J, Mengual E, et al. Artificial intelligence in undergraduate medical education: an updated scoping review. *BMC Med Educ*. 2025;25(1):1609. doi:10.1186/s12909-025-08188-2
54. Zhu L, Lai Y, Mou W, et al. ChatGPT's ability to generate realistic experimental images poses a new challenge to academic integrity. *J Hematol Oncol*. 2024;17(1):27. doi:10.1186/s13045-024-01543-8
55. Li D, Lebai Lutfi S. Large Language Model–Based Virtual Patient Systems for History-Taking in Medical Education: comprehensive Systematic Review. *JMIR Med Inform*. 2026;14:e79039. doi:10.2196/79039
56. Sadeq MA, Ghorab RMF, Ashry MH, et al. AI chatbots show promise but limitations on UK medical exam questions: a comparative performance study. *Sci Rep*. 2024;14(1):18859. doi:10.1038/s41598-024-68996-2
57. Rahimzadeh V, Kostick-Quenet K, Blumenthal Barby J, et al. Ethics Education for Healthcare Professionals in the Era of ChatGPT and Other Large Language Models: do We Still Need It? *The American Journal of Bioethics*. 2023;23(10):17–27. doi:10.1080/15265161.2023.2233358
58. Tailor PD, D'Souza HS, Castillejo Becerra CM, et al. Evaluation of AI Summaries on Interdisciplinary Understanding of Ophthalmology Notes. *JAMA Ophthalmol*. 2025;143(5):410. doi:10.1001/jamaophthalmol.2025.0351
59. Akgün FE, Akgün M. Governing Generative AI in Healthcare: a Normative Conceptual Framework for Epistemic Authority, Trust, and the Architecture of Responsibility. *Healthcare*. 2026;14(8):1098. doi:10.3390/healthcare14081098
60. Johri S, Jeong J, Tran BA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat Med*. 2025;31 (1):77–86. doi:10.1038/s41591-024-03328-5

61. Wang H, Shan W, Liu R, et al. Can large language models serve as digital assistants for medical undergraduates? – a bibliometric mapping and scoping analysis of the medical-education literature. *DIGITAL HEALTH*. 2025;11:20552076251390280. doi:10.1177/20552076251390280
62. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare*. 2023;11(6):887. doi:10.3390/healthcare11060887
63. Shool S, Adimi S, Saboori Amleshi R, et al. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inform Decis Mak*. 2025;25(1):117. doi:10.1186/s12911-025-02954-4
64. Zhang Q, Huang Z, Huang Y, et al. Generative AI in medical education: feasibility and educational value of LLM-generated clinical cases with MCQs. *BMC Med Educ*. 2025;25:1502. doi:10.1186/s12909-025-08085-8
65. Small WR, Austrian J, O'Donnell L, et al. Evaluating hospital course summarization by an electronic health record–based large language model. *JAMA Netw Open*. 2025;8(8):e2526339. doi:10.1001/jamanetworkopen.2025.26339
66. Knott M, Krebs M, Kerscher A. Large language models in healthcare quality management: a European perspective on process automation and compliance. *Front Digit Health*. 2026;8:1761641. doi:10.3389/fdgth.2026.1761641
67. Alkhaaldi SMI, Kassab CH, Dimassi Z, et al. Medical student experiences and perceptions of chatgpt and artificial intelligence: cross-sectional study. *JMIR Med Educ*. 2023;9:e51302. doi:10.2196/51302
68. Li X, Yan X, Lai H. The ethical challenges in the integration of artificial intelligence and large language models in medical education: a scoping review. *PLoS One*. 2025;20:e0333411. doi:10.1371/journal.pone.0333411
69. Soares VV, Passos FFDC, Da Silva ICS, et al. The Role of Large Language Models in Teaching Psychiatric Semiology: a Systematic Review. *Trends Psychiatry Psychother*. 2026. doi:10.47626/2237-6089-2025-1237
70. Vrdoljak J, Boban Z, Vilović M, et al. A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. *Healthcare*. 2025;13(6):603. doi:10.3390/healthcare13060603
71. Zhong X, Li S, Chen Z, et al. Considerations for Patient Privacy of Large Language Models in Health Care: scoping Review. *J Med Internet Res*. 2025;27:e76571. doi:10.2196/76571
72. Li Z, Fu Q, Zhao Y, et al. Empowering standardized residency training in China through large language models: problem analysis and solutions. *Annals of Medicine*. 2025;57(1):2516695. doi:10.1080/07853890.2025.2516695
73. Hou J, An F, Qin H, et al. Application of DeepSeek-assisted problem-based learning in hematology residency training. *BMC Med Educ*. 2025;25 (1):1291. doi:10.1186/s12909-025-07852-x
74. Bahmani A, Cha K, Alavi A, et al. Achieving inclusive healthcare through integrating education and research with AI and personalized curricula. *Commun Med*. 2025;5(1):356. doi:10.1038/s43856-025-01034-y
75. Divito CB, Katchikian BM, Gruenwald JE, et al. The tools of the future are the challenges of today: the use of ChatGPT in problem-based learning medical education. *Medical Teacher*. 2024;46(3):320–322. doi:10.1080/0142159X.2023.2290997
76. Nguyen T. ChatGPT in Medical Education: a Precursor for Automation Bias? *JMIR Med Educ*. 2024;10:e50174. doi:10.2196/50174
77. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: systematic Review and Meta-Analysis. *J Med Internet Res*. 2024;26:e60807. doi:10.2196/60807
78. Jowsey T, Stokes-Parish J, Singleton R, et al. Medical education empowered by generative artificial intelligence large language models. *Trends in Molecular Medicine*. 2023;29(12):971–973. doi:10.1016/j.molmed.2023.08.012