



GENERATIVE ARTIFICIAL INTELLIGENCE AND LARGE LANGUAGE MODELS AS METHODOLOGICAL SUPPORT TOOLS IN HEALTH RESEARCH: A BIBLIOMETRIC ANALYSIS

INTELIGÊNCIA ARTIFICIAL GENERATIVA E MODELOS DE LINGUAGEM DE GRANDE ESCALA COMO FERRAMENTAS DE APOIO METODOLÓGICO NA PESQUISA EM SAÚDE: UMA ANÁLISE BIBLIOMÉTRICA

INTELENCIA ARTIFICIAL GENERATIVA Y MODELOS DE LENGUAJE DE GRAN TAMAÑO COMO HERRAMIENTAS DE APOYO METODOLÓGICO EN LA INVESTIGACIÓN EN SALUD: UN ANÁLISIS BIBLIOMÉTRICO

Yasmin Saba de Almeida¹

Elaine Antunes Cortez¹

Geilsa Soraia Cavalcanti Valente¹

ORCID: 0000-0003-2391-7009

ORCID: 0000-0003-3912-9648

ORCID: 0000-0003-4488-4912

¹ Universidade Federal Fluminense, Escola de Enfermagem Aurora de Afonso Costa. Niterói, Rio de Janeiro, Brazil.

How to cite: Almeida YS, Cortez EA, Valente GSC. Generative Artificial Intelligence and Large Language Models as methodological support tools in health research: a bibliometric analysis. Online Braz J Nurs. 2026;25(Suppl1):e20267204. <https://doi.org/10.17665/1676-4285.20267204>

RESUMO

Objetivo: Mapear a produção científica e as tendências temáticas do uso de inteligência artificial generativa (GenAI) e modelos de linguagem de grande escala (LLMs) como ferramentas de apoio metodológico na pesquisa em saúde. **Método:** Estudo bibliométrico realizado em setembro de 2026 nas bases PubMed, Scopus e Web of Science. Os registros foram triados por meio de um pipeline híbrido (humano-IA) com apoio de um LLM segundo critérios previamente definidos e revisão humana. Os dados foram analisados no RStudio com uso do pacote *bibliometrix* e de sua interface *biblioshiny*. **Resultados:** O corpus final compreendeu 552 documentos, distribuídos em 337 fontes, com idade média de 0,88 anos e taxa anual de crescimento de 70,7%. Os Estados Unidos concentraram o maior número de afiliações de autores, seguidos pela China. A literatura apresentou alta dispersão entre periódicos e forte concentração em autores ocasionais (89,2% com apenas uma publicação), além de modesta colaboração internacional (14,7%). O mapa temático revelou que os temas motores permanecem centrados no uso genérico de LLMs e ChatGPT, enquanto aplicações metodológicas específicas ainda não constituem agrupamentos consolidados. **Conclusão:** O uso da GenAI e dos LLMs como apoio metodológico à pesquisa em saúde constitui um campo emergente e em rápida expansão, mas ainda desprovido de uma estrutura temática e de canais de publicação consolidados. A consolidação da área exige o estabelecimento de critérios normativos de validação, transparência no relato, supervisão humana estrita e letramento dos pesquisadores em inteligência artificial, garantindo a integridade das evidências científicas produzidas, com repercussões diretas na qualificação das práticas e na segurança do cuidado em saúde.

Descritores: Bibliometria; Inteligência Artificial Generativa; Modelos de Linguagem de Grande Escala; Saúde; Técnicas de Pesquisa.

ABSTRACT

Objective: To map the scientific output and thematic trends in the use of generative artificial intelligence (GenAI) and large language models (LLMs) as methodological support tools in health research. **Method:** Bibliometric study conducted in September 2026 in the PubMed, Scopus, and Web of Science databases. Records were screened through a hybrid (human-AI) pipeline, in which an LLM applied predefined criteria and a human reviewer checked the results. Data were analyzed in RStudio using the *bibliometrix* package and its *biblioshiny* interface. **Results:** The final corpus comprised 552 documents from 337 sources, with a mean document age of 0.88 years and an annual growth rate of 70.7%. The United States accounted for the largest number of author affiliations, followed by China. The literature showed high dispersion across journals and a strong concentration of occasional authors (89.2% with a single publication), along with modest international collaboration (14.7%). The thematic map revealed that motor themes remain centered on the generic use of LLMs and ChatGPT, whereas specific methodological applications have not yet formed consolidated clusters. **Conclusion:** The use of GenAI and LLMs as methodological support in health research is an emerging and rapidly expanding field, but it still lacks a consolidated thematic structure and established publication venues. Consolidating the field will require normative validation criteria, transparent reporting, strict human oversight, and AI literacy among researchers, so as to safeguard the integrity of the scientific evidence produced, with direct implications for the quality of practice and the safety of health care.

Descriptors: Bibliometrics; Generative Artificial Intelligence; Large Language Models; Health; Investigative Techniques.

RESUMEN

Objetivo: Mapear la producción científica y las tendencias temáticas del uso de la inteligencia artificial generativa (GenAI) y de los modelos de lenguaje de gran tamaño (LLM) como herramientas de apoyo metodológico en la investigación en salud. **Método:** Estudio bibliométrico realizado en septiembre de 2026 en las bases de datos PubMed, Scopus y Web of Science. Los registros se cribaron mediante un flujo híbrido (humano-IA) en el que un LLM aplicó criterios definidos previamente y una revisora humana verificó los resultados. Los datos se analizaron en RStudio con el paquete *bibliometrix* y su interfaz *biblioshiny*. **Resultados:** El corpus final estuvo compuesto por 552 documentos, distribuidos en 337 fuentes, con una antigüedad media de 0,88 años y una tasa de crecimiento anual del 70,7%. Estados Unidos concentró el mayor número de afiliaciones de autores, seguido de China. La literatura mostró una alta dispersión entre revistas y una fuerte concentración de autores ocasionales (el 89,2% con una sola publicación), además de una colaboración internacional modesta (14,7%). El mapa temático reveló que los temas motores siguen centrados en el uso genérico de los LLM y de ChatGPT, mientras que las aplicaciones metodológicas específicas aún no conforman agrupaciones consolidadas. **Conclusión:** El uso de la GenAI y de los LLM como apoyo metodológico a la investigación en salud constituye un campo emergente y en rápida expansión, pero que todavía carece de una estructura temática y de canales de publicación consolidados. La consolidación del área exige establecer criterios normativos de validación, transparencia en la comunicación de los resultados, una supervisión humana estricta y la alfabetización de los investigadores en inteligencia artificial, de modo que se garantice la integridad de la evidencia científica producida, con repercusiones directas en la calidad de las prácticas y en la seguridad de la atención en salud.

Descriptores: Bibliometría; Inteligencia Artificial Generativa; Modelos de Lenguaje de Gran Escala; Salud; Técnicas de Investigación.

Editors:

Rosimere Ferreira Santana (ORCID: 0000-0002-4593-3715)

Audrey Vidal Pereira (ORCID: 0000-0002-6570-9016)

Publisher:

Escola de Enfermagem Aurora de Afonso Costa – UFF

Rua Dr. Celestino, 74 – Centro, CEP: 24020-091 – Niterói, RJ, Brazil

Journal email: objn.cme@id.uff.br

Corresponding author:

Yasmin Saba de Almeida

E-mail: yasminsaba@id.uff.br

What is already known:

- Generative artificial intelligence and large language models have expanded the use of artificial intelligence in activities involving the creation, analysis, and organization of knowledge.
- In health, these technologies are used to support diagnosis, education, academic writing, and the discovery and synthesis of scientific literature.
- The use of generative artificial intelligence as methodological support across the stages of health research remains insufficiently explored in a systematic manner.

What this article adds:

- Presents a specific bibliometric overview of the use of generative artificial intelligence as methodological support in health research.
- Identifies the leading authors, journals, countries, collaboration networks, and thematic trends in this field.
- Reveals a recent and rapidly expanding field and indicates priorities for its scientific and methodological consolidation.

INTRODUCTION

Generative artificial intelligence (GenAI), particularly large language models (LLMs), has rapidly evolved from a writing assistant into a tool applied in domains of high epistemic complexity, including several core stages of the scientific research process⁽¹⁾. In the biomedical ecosystem, this paradigm shift has been accompanied by the expansion of artificial intelligence (AI)-assisted tools for literature discovery and exploration, ranging from text-mining platforms to relational and semantic knowledge representation^(2,3).

In health research, evidence synthesis has emerged as one of the main domains of application for GenAI and LLMs. These models are increasingly being incorporated into the development of search strategies, automated or semi-automated literature screening, parameterized data extraction, and the appraisal of methodological quality or risk of bias. However, evidence synthesized in a recent scoping review cautions that the technological maturity of these tools does not yet guarantee their reproducibility or accuracy without critical oversight⁽⁴⁾.

Although the volume of publications on AI in health has grown exponentially, studies devoted to strictly methodological applications within the research workflow remain diluted among studies focused on diagnostic support, health education, and academic writing assistance. Preliminary searches in PubMed, Scopus, and Web of Science confirmed the absence of bibliometric studies that map, with scientific rigor and systematic detail, the use of these technologies as methodological support in health research. This gap matters because health research depends on the timely and rigorous production of evidence to inform care, guideline development, and health policy. By supporting labor-intensive stages such as searching, screening, data extraction, and synthesis, GenAI and LLMs have great potential to streamline the evidence production cycle, provided that their technical validation, accuracy, and procedural reproducibility are properly ensured^(4,5).

In this sense, understanding how this field has been taking shape makes it possible to identify which stages of health research already concentrate efforts, which authors, institutions, and sources are driving scientific output, and which gaps remain. This overview can guide research priorities and promote the use of GenAI and LLMs in ways that are more consistent with the needs and rigor of health research. Bibliometric analysis allows researchers to examine the growth of scientific output, collaboration networks, and the themes that structure the knowledge

produced, offering an accurate picture of the field's stage of maturity^(6,7). Accordingly, this study aims to map the scientific output and thematic trends in the use of GenAI and LLMs as methodological support tools in health research.

METHOD

This is a scientometric and bibliometric study, characterized by the application of quantitative methods to analyze scientific output in a specific field of knowledge. Bibliometrics uses statistical indicators to examine patterns of publication, citation, and scientific collaboration, making it possible to identify knowledge structures, productivity dynamics, and the thematic domains of a discipline^(6,7).

Research question

The guiding question of the study was: What is the bibliometric profile, and what are the thematic trends, of the scientific literature on GenAI and LLMs used as methodological support tools in health research?

Eligibility criteria

We included documents in which GenAI, especially LLMs, was used, evaluated, or discussed specifically as support for at least one methodological stage of health research. Eligible stages included study conception and methodological design, search strategy development, screening of titles, abstracts, or full texts, data extraction and synthesis, appraisal of methodological quality or risk of bias, quantitative or qualitative data analysis, and research hypothesis generation.

We excluded documents that addressed these technologies exclusively in the context of clinical practice or patient care, health education or training, scientific writing or manuscript editing, or that merely mentioned AI without an identifiable methodological application. Studies outside the health field were also excluded.

No restrictions were applied regarding language, document type, or publication period.

Literature search

Searches were conducted on September 1, 2026, in the indexed databases PubMed, Scopus, and Web of Science. The strategies combined the controlled vocabulary of *Medical Subject Headings* (MeSH) with free-text terms related to GenAI and LLMs, the research process, and

health research. Boolean operators and exclusion terms were also used to reduce the retrieval of studies focused exclusively on clinical applications, health education, or scientific writing. The complete search strategies are available for audit in the project's Open Science Framework (OSF) repository, under DOI: 10.17605/OSF.IO/NC8K9.

GenAI-assisted study selection

A computational workflow was developed in Python to automate and log deduplication, record classification, and screening decisions. The workflow was built with the support of the GenAI tools Codex, by OpenAI, and Claude Code, by Anthropic, using Visual Studio Code, version 1.137.0, as the development environment. The eligibility criteria and the structured *prompt* were defined and iteratively refined by the researcher throughout this process. The source code and the version of the *prompt* used for classification are available in the public GitHub repository [enfermagempesquisadora/genai-health-research-screening](https://github.com/enfermagempesquisadora/genai-health-research-screening).

After deduplication by exact and semantic matching, titles and abstracts were processed with the support of the LLM Claude. For each record, the model indicated whether to include, exclude, or classify it as uncertain, and also recorded the exclusion category (when applicable), a brief justification, an excerpt from the title or abstract supporting the decision, and its self-reported confidence level. Supporting excerpts were automatically checked against the source text. The researcher reviewed all classifications and prioritized the assessment of records classified as uncertain, making the final screening decisions under full human oversight.

Data processing and analysis

Data were extracted, harmonized, and analyzed in the R environment (via the RStudio interface), using the *bibliometrix*⁽⁶⁾ package and its graphical interface, *biblioshiny*. *Bibliometrix* is an open-source tool for bibliometric analysis, with advanced features for assessing productivity, collaboration, citation, and science mapping⁽⁶⁾. Complementary descriptive analyses were performed in Microsoft Excel spreadsheets.

The informetric analysis was structured around the three classic laws of bibliometrics: 1) Bradford's Law (which assessed the dispersion of articles across publication sources, allowing the empirical delimitation of the core journals of the corpus)⁽⁸⁾; 2) Lotka's Law (which examined the distribution of author productivity, testing the statistical fit between the observed distribution and the theoretical power-law model using the Kolmogorov-Smirnov test)⁽⁹⁾; and 3) Zipf's Law (which described the frequency distribution of terms, underpinning the keyword co-occurrence analyses and the conceptual structure of the field)⁽¹⁰⁾.

We also analyzed annual output, annual growth rate, document type and language, mean publication age, geographic distribution of author affiliations, most represented institutions, most productive and most cited countries, most productive or most cited sources and authors, keyword co-occurrence, and thematic mapping. Output and

impact indicators included number of publications, total citations, and citations per document. Citations were taken as recorded in the metadata of the imported records and interpreted as descriptive measures of the corpus, given the differences in coverage across databases⁽¹¹⁾.

The strategic diagram positioned the thematic clusters in a two-dimensional space according to centrality (degree of relevance and inter-cluster connectivity) and density (degree of internal development of the cluster). Themes were classified into four quadrants: motor themes, niche themes, emerging or declining themes, and basic or transversal themes.

Ethical aspects

Because the study used only bibliographic records and publicly available information, with no identifiable participant data, it did not require review by the CEP/Conep System, in accordance with National Health Council (CNS) Resolution No. 510/2016⁽¹²⁾. GenAI tools were used to support the development of the analysis workflow, the screening of records, and the writing of the manuscript, under the researcher's supervision and the direct intellectual responsibility of the authors.

RESULTS

Selection flow and corpus eligibility

A total of 2,707 records were identified in the selected databases. After removal of 1,116 duplicates (1,042 identified by DOI and 74 by title and publication year), 1,591 unique records underwent title and abstract screening. The hybrid (human-AI) screening pipeline yielded 552 records eligible for the bibliometric corpus and 938 exclusions. Of the excluded records, 587 were off-topic, 287 addressed clinical applications, 34 dealt with scientific writing, 29 with health education, and one had no identifiable methodological application of its own.

The 101 records initially classified as uncertain were excluded from the final corpus after qualitative assessment, as it could not be confirmed that GenAI or LLMs were used or evaluated in, or supported, a methodological stage of health research. These also included off-topic studies, studies focused on clinical applications, scientific writing, or health education, and documents with no methodological content of their own. Figure 1 shows the study selection flow.

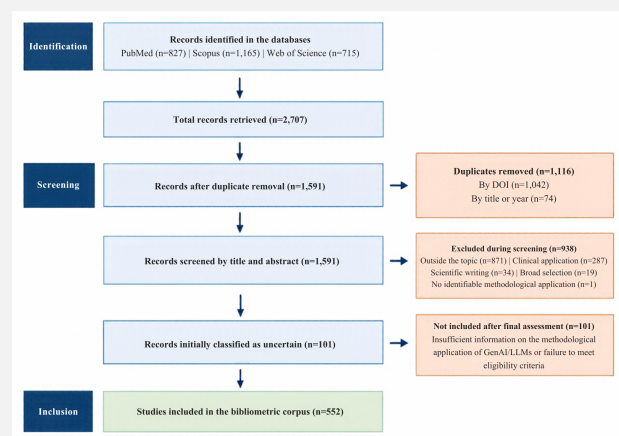
Characteristics and evolution of scientific output

The 552 documents were distributed across 337 sources. Original journal articles predominated (n=425; 77.0% of the corpus), followed by conference papers (n=68), reviews (n=38), letters (n=13), editorials (n=3), book chapters (n=3), and notes (n=2). English was the predominant language, present in 98.7% of the records. The mean number of citations per document was 2.62.

Scientific output spanned the publication years 2016 to 2027 (Figure 2). One document was identified in 2016, with no records between 2017 and 2022. From 2023 onward, output increased markedly, with 19 documents in

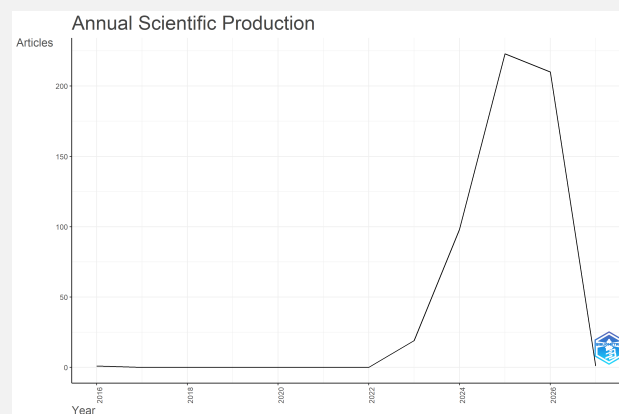
2023, 98 in 2024, 223 in 2025, and 210 up to September 1, 2026. The only document recorded for 2027 was a conference paper already available online in 2026 but assigned to a 2027 volume. This record was kept in the corpus and was not interpreted as a trend in output after the search cutoff date. The mean document age was 0.88 years, underscoring how recent the field is.

Figure 1 – Study selection flowchart. Niterói, RJ, Brazil, 2026



Source: prepared by the authors, 2026.

Figure 2 – Annual scientific output on GenAI and LLMs as methodological support tools in health research, 2016–2027 (n=552). Niterói, RJ, Brazil, 2026



Note: The only document with a 2027 publication year was available online in 2026 and was kept in the corpus because it was retrieved during the search period.

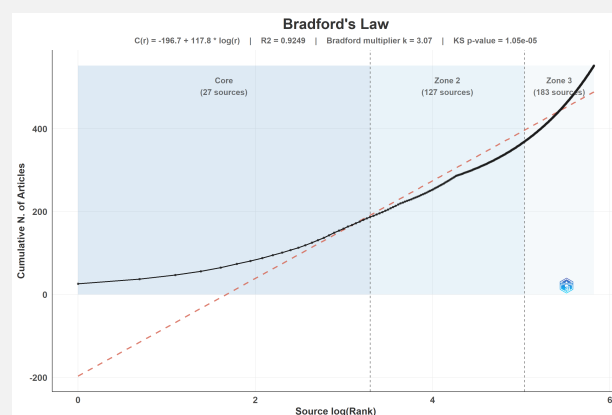
Source: research data, generated in Biblioshiny, 2026.

Dispersion of publication sources and citation impact (Bradford's Law)

The distribution of the 337 sources into Bradford zones identified a core of 27 sources, accounting for 187 documents, or 33.9% of the corpus (Figure 3). The second zone comprised 127 sources and 182 documents, while

the third included 183 sources and 183 documents. The Bradford model showed an excellent linear fit ($R^2 = 0.9249$), with a multiplier of 3.07. The Kolmogorov-Smirnov test confirmed a statistical discontinuity between the observed empirical distribution and the ideal theoretical model ($D = 0.1899$; $p < 0.001$).

Figure 3 – Distribution of sources across Bradford zones. Niterói, RJ, Brazil, 2026



Source: research data, generated in Biblioshiny, 2026.

The journal *Studies in Health Technology and Informatics* stood out as the most productive source (n=26). In terms of citation impact, the *Journal of Clinical Epidemiology* led the corpus, with 149 citations accumulated across six publications (Table 1).

Among the articles with the highest global citation counts, those focused on semantic exploration and evidence synthesis stood out. The 2024 article by Wei CH on the PubTator 3.0 platform was the most cited in the corpus (n=104), followed by the studies by Lieberum JL (2025; n=103), Aljamaan F (2024; n=97), Wang S (2024; n=92), and Jin Q (2024; n=91).

Authorship, productivity, and collaboration patterns (Lotka's Law)

The corpus included 2,914 authors. Wang Y was the most productive author (n=18), followed by Wang Z (n=13), Chen Y (n=12), and Li Y (n=12). Output was predominantly collaborative, with a mean of 6.32 co-authors per article. Only 21 documents (3.8%) were single-authored. The international collaboration rate was 14.7%.

Applying Lotka's Law showed that 2,599 authors (89.2% of the total) published only one article in the field (Table 2). The empirical distribution differed from the classic theoretical Lotka model, with exponent $\beta = 2.0$ ($D = 0.6429$; $p = 0.006$). However, the power-law model fitted to the corpus yielded $\beta = 2.87$ (constant $C = 0.444$; $R^2 = 0.928$) and did not differ significantly from the observed data ($D = 0.2857$; $p = 0.617$). This high β exponent reflects an atypical concentration of occasional authors, characteristic of emerging fields.

Table 1 – Main sources in the corpus by number of publications and available citations. Niterói, RJ, Brazil, 2026

Source	NP	TC	CPP (TC/NP)	% of total (NP/552 × 100)
Studies in Health Technology and Informatics	26	11	0.42	4.71
Journal of Medical Internet Research	11	0	0.00	1.99
bioRxiv: The Preprint Server for Biology	10	0	0.00	1.81
JMIR AI	9	0	0.00	1.63
Journal of the American Medical Informatics Association	9	0	0.00	1.63
Scientific Reports	9	0	0.00	1.63
Bioinformatics (Oxford, England)	7	0	0.00	1.27
BMC Medical Research Methodology	7	33	4.71	1.27
Cureus	7	0	0.00	1.27
BMJ Open	6	1	0.17	1.09

Note: NP = number of publications; TC = total citations; CPP = citations per publication.

Source: research data, 2026.

Table 2 – Distribution of authors by number of published documents and theoretical proportion under the classic Lotka's Law. Niterói, RJ, Brazil, 2026

Number of publications	Number of authors	Observed proportion (%)	Theoretical proportion (%)
1	2599	89.2	60.8
2	210	7.2	15.2
3	60	2.1	6.8
4	15	0.5	3.8
5	10	0.3	2.4
6	5	0.2	1.7
7	5	0.2	1.2
8	1	0.0	0.9
9	1	0.0	0.8
10	3	0.1	0.6
11	1	0.0	0.5
12	2	0.1	0.4
13	1	0.0	0.4
18	1	0.0	0.2

Note: The theoretical proportion corresponds to the classic Lotka's Law model, calculated as $f(n) = 0.6079/n^2$.

Source: research data, 2026.

Geographically, the United States led in the frequency of affiliations (881 occurrences), followed by China (502 occurrences). The most frequent institution was Lanzhou University (n=102), followed by the National Institute on Aging (n=47), University of California (n=44), McMaster University (n=37), and University of Pennsylvania (n=35). In terms of citations by country of affiliation, the United States accumulated 228 citations (mean of 9.9 citations per article), whereas China totaled 14 citations (mean of 2.0 per article). Jin Q had the highest number of citations among authors, with 195, while Wang Y stood out for productivity.

Conceptual structure and thematic map

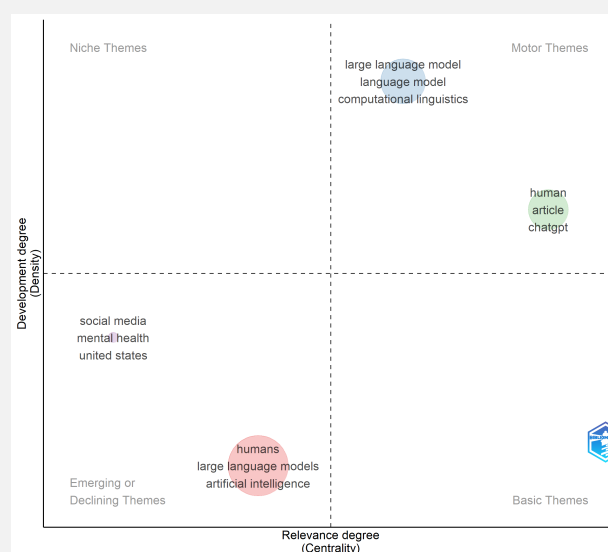
The two-dimensional thematic map (Figure 4), built from the authors' keywords, organized the conceptual structure into four clusters, as follows:

Motor Themes (upper-right quadrant): two well-defined clusters were identified. The first grouped the terms “*large language model*”, “*language model*”, and “*computational linguistics*”; the second brought together “*ChatGPT*”, “*human*”, and “*article*”, indicating the strong development and methodological centrality of LLMs in the research ecosystem.

Emerging or Declining Themes (lower-left quadrant): this quadrant grouped the terms “*humans*”, “*large language models*”, and “*artificial intelligence*”, as well as a cluster formed by “*social media*”, “*mental health*”, and “*United States*”.

Niche and Basic Themes: No structured clusters were identified in these quadrants, reinforcing the concentrated and still dynamic nature of the literature analyzed.

Figure 4 – Thematic map of author keywords associated with GenAI and LLMs as methodological support tools in health research. Niterói, RJ, Brazil, 2026



Source: research data, generated in Biblioshiny, 2026.

DISCUSSION

Growth dynamics and stage of consolidation of the field

Compared with broader bibliometric analyses of GenAI in health, the literature focused strictly on methodological support for research is still more circumscribed and at an early stage of consolidation. A bibliometric analysis of ChatGPT in health, which included clinical, educational, and patient communication applications, gathered 594 publications and recorded an annual growth of 91.18% between 2023 and 2024⁽¹³⁾. Another analysis, focused on the use of the same LLM in education in Asia, identified 1,665 documents published between 2022 and 2025, with an annual growth rate of 545.53%⁽¹⁴⁾. In the present study, this rate was 70.7%, lower than those observed in both analyses. Differences in databases, time frames, and search strategies preclude a direct comparison of these results. Even so, the findings show that the use of GenAI and LLMs as methodological support in health research is also expanding, albeit less steeply and with greater rigor in the selection of use cases.

Informetrics of authorship networks, sources, and transnational collaboration

Taken together with the collaboration indicators, the analysis of Bradford's and Lotka's Laws points in the same direction. Although some journals already concentrate a substantial share of the publications, the documents are spread across 337 sources, indicating that the topic does not yet have fully established dissemination venues. Author productivity was also heavily concentrated in one-off contributions, with 89.2% of authors appearing in only one document, reflecting a high proportion of transient authors, and productivity dropped more steeply than predicted by the classic Lotka model ($\beta = 2.87$). International collaboration, present in 14.7% of the documents, completes this picture by showing that cross-country connections are still infrequent. Overall, these results suggest that the growth of the literature has been driven more by the entry of new authors and the spread of the topic across different journals than by the formation of a stable, consolidated research community.

Geographic hegemony and implications for scientific equity

The concentration of affiliations in the United States and China should be interpreted in light of cross-country differences in infrastructure, governance, workforce, and institutional capacity for adopting AI in health⁽¹⁵⁾. Although this study did not directly assess these inequalities, the results show that a substantial share of the knowledge on the use of GenAI in health research is being produced in countries with high scientific and technological capacity. This predominance may shape the applications investigated, the criteria used to evaluate them, and the problems considered priorities. Technologies developed and validated predominantly in these settings may therefore not respond equally well to the needs of health systems and populations operating under resource constraints or with distinct

epidemiological profiles.

Evolution of thematic structure and maturity

The thematic map showed that the most developed and connected themes remain centered on the generic evaluation of LLMs, with ChatGPT standing out. This indicates that the literature is still organized around the technologies themselves and their potential uses, whereas more specific methodological applications (such as automated evidence synthesis or risk-of-bias modeling) have not yet emerged as consolidated themes. The absence of basic or niche themes and the classification of the remaining clusters as emerging or declining reinforce this picture. Thus, although GenAI is already being incorporated into different stages of health research, no well-defined thematic structure around specific methodological uses can yet be observed.

Methodological uncertainties, integrity, and reproducibility

The literature already describes the use of GenAI in searching, screening, data extraction, risk-of-bias assessment, and evidence synthesis. However, consistent standardized criteria are still lacking to define at which research stages LLMs can be used safely, what error levels are acceptable, and how their performance should be audited and validated^(4,5). In health research, this uncertainty is especially relevant because errors in identifying, selecting, or interpreting evidence can compromise reviews that inform protocols, guidelines, policies, and care decisions, with potential repercussions for the quality and safety of care⁽⁵⁾.

For this reason, the incorporation of GenAI into health research should be guided not only by its potential to speed up stages of the work but also by its capacity to preserve the rigor, reliability, and reproducibility of the evidence produced. Reports of its use should include the tool and its version, the purpose of use, the instructions employed, the validation methods, the verification of outputs, and the privacy aspects involved⁽¹⁶⁾. This level of detail helps clarify the benefits and limitations of the technology in the context in which it was used and makes the decisions taken throughout the process auditable. Even when these tools are used, the planning, validation, and interpretation of results remain the sole responsibility of the researchers⁽¹⁶⁾.

AI literacy and the transformation of the researcher's professional identity

Finally, the expanding use of GenAI and LLMs in the research process brings new challenges for researcher training. Their use requires not only operational command of the tools but also AI literacy and the ability to critically appraise generated outputs, recognize algorithmic limitations, and justify methodological decisions. A systematic review of AI education and training for health workers pointed to the need for training initiatives covering technical skills, critical appraisal, ethical aspects, and the safe use of these technologies⁽¹⁷⁾, while another study showed that professional identity may be related to AI readiness among health workers⁽¹⁸⁾. These findings suggest that researcher training needs to keep pace with these changes, addressing

not only the competencies required to use GenAI but also an understanding of researchers' role and responsibility in the face of new ways of producing knowledge^(17,18).

Study limitations

As limitations, the results reflect the databases consulted, the search strategy, and the time frame adopted. Citation indicators were obtained from the imported meta-data and should be interpreted as descriptive measures of the corpus, since coverage varies across sources. The use of an LLM to support screening should also be considered, as automated classifications are prone to errors and may have influenced the composition of the corpus despite human oversight. At the same time, using it in the selection process itself made it possible to explore, in practice, one of the methodological applications that are the object of this study. Moreover, because the literature on this topic is growing rapidly, the overview presented is bound to the period analyzed and may change as new studies are published.

CONCLUSION

This study showed that the use of GenAI and LLMs as methodological support in health research is a recent and expanding field, yet one still marked by the dispersion of publications across authors and journals, low international collaboration, and the absence of a consolidated thematic structure around specific methodological applications. The findings indicate that the adoption of these technologies is advancing faster than the development of a scientific and regulatory foundation capable of guiding their use in health research.

We recommend that future studies validate the performance of these tools in specific methodological stages, that their use be reported in a standardized and transparent manner, with human oversight, and that AI literacy be

incorporated into researcher training, including in nursing. We also recommend strengthening collaborative networks, expanding research in Global South countries, which tend to adopt AI technologies developed in other contexts, and periodically updating this mapping. These measures are necessary because errors in searching, selecting, analyzing, and synthesizing evidence can affect the quality and safety of care, as well as health guidelines and policies.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

FUNDING

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001.

ARTIFICIAL INTELLIGENCE USE STATEMENT

GenAI tools were used at different stages of this study. Codex and Claude Code were used to develop the Python computational workflow. Title and abstract screening was performed with the support of Claude Opus 5, based on eligibility criteria and a prompt defined and refined by the researcher. ChatGPT and Claude were also used to support the organization, writing, and language editing of the manuscript.

These tools were used under full human oversight. The classifications generated by Claude during screening were reviewed by the researcher, with priority given to records classified as uncertain and decisions adjusted when necessary. The definition of eligibility criteria, the methodological decisions, the interpretation of results, and the conclusions remained the responsibility of the researchers, who critically reviewed the final version and took full responsibility for its content.

REFERENCES

1. Truhn D, Reis-Filho JS, Kather JN. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nat Med*. 2023;29(12):2983-2984. <https://doi.org/10.1038/s41591-023-02594-z>. PMID: 37853138.
2. Wei C, Allot A, Lai P, Leaman R, Tian S, Luo L, et al. PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Research*. 2024;52(W1):W540-W546. <https://doi.org/10.1093/nar/gkae235>. PMID: 38572754.
3. Jin Q, Leaman R, Lu Z. PubMed and beyond: biomedical literature search in the age of artificial intelligence. *eBioMedicine*. 2024;100:104988. <https://doi.org/10.1016/j.ebiom.2024.104988>. PMID: 38306900.
4. Lieberum J, Toews M, Metzendorf M, Heilmeyer F, Siemens W, Haverkamp C, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use—a scoping review. *Journal of Clinical Epidemiology*. 2025;181:111746. <https://doi.org/10.1016/j.jclinepi.2025.111746>. PMID: 40021099.
5. Siemens W, von Elm E, Binder H, Böhringer D, Eisele-Metzger A, Gartlehner G, et al. Opportunities, challenges and risks of using artificial intelligence for evidence synthesis. *BMJ EBM*. 2025;30(6):381-384. <https://doi.org/10.1136/bmjebm-2024-113320>. PMID: 39788693.
6. Aria M, Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*. 2017;11(4):959-975. <https://doi.org/10.1016/j.joi.2017.08.007>.
7. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*. 2021;133:285-296. <https://doi.org/10.1016/j.jbusres.2021.04.070>.
8. Bradford SC. Sources of information on specific subjects. *Engineering*. 1934;137:85-86.
9. Lotka AJ. The frequency distribution of scientific productivity. *J Wash Acad Sci*. [Internet]. 1926 [cited 2026 Sep 12];16(12):317-23. Available from: <https://www.jstor.org/stable/24529203>.
10. Zipf GK. Human behavior and the principle of least effort: an introduction to human ecology. Cambridge: Addison-Wesley; 1949.
11. Gerasimov I, KC B, Mehrabian A, Acker J, McGuire MP. Comparison of datasets citation coverage in Google Scholar, Web of Science, Scopus, Crossref, and DataCite. *Scientometrics*. 2024;129(7):3681-3704. <https://doi.org/10.1007/s11192-024-05073-5>.
12. Brasil. Conselho Nacional de Saúde. Resolução nº 510, de 7 de abril de 2016 [Resolution No. 510 of April 7, 2016].

Provides for the rules applicable to research in the Human and Social Sciences whose methodological procedures involve the use of data obtained directly from participants or of identifiable information, or that may entail risks greater than those of everyday life, as defined in this Resolution] [Internet]. Brasília: Ministério da Saúde; 2016 [cited 2026 Sep 12]. Available from: <https://www.gov.br/conselho-nacional-de-saude/pt-br/atos-normativos/resolucoes/2016/resolucao-no-510.pdf/view>.

13. Amir AS. Revolutionizing healthcare with ChatGPT: current advancements, challenges, and future directions. *Discov Artif Intell.* 2026;6(1):582. <https://doi.org/10.1007/s44163-026-01506-7>.
14. Irwanto I. Research on ChatGPT in education in Asia: a bibliometric analysis (2022–2025). *Discov Artif Intell.* 2025;5(1):244. <https://doi.org/10.1007/s44163-025-00505-4>.
15. World Health Organization, Regional Office for Europe. Artificial intelligence is reshaping health systems: state of readiness across the European Union [Internet]. Copenhagen: WHO Regional Office for Europe; 2026 [cited 2026 Sep 12]. Available from: <https://www.who.int/europe/publications/item/WHO-EURO-2026-12707-52481-81471>
16. Luo X, Tham YC, Giuffrè M, Ranisch R, Daher M, Lam K, et al. Reporting guideline for the use of Generative Artificial intelligence tools in MEDical Research: the GAMER Statement. *BMJ EBM.* 2025;30(6):390-400. <https://doi.org/10.1136/bmjebm-2025-113825>. PMID: 40360239.
17. Woods L, Lyons K, Van Der Vegt A, Olsen Q, Huang W, Khor JS, et al. Assessing the effectiveness of artificial intelligence education and training for healthcare workers: a systematic review. *BMC Med Educ.* 2026;26(1). <https://doi.org/10.1186/s12909-026-08969-3>. PMID: 41807961.
18. Ta'an W, Damrah S, Al-Hammouri MM, Williams B. Professional identity and its relationships with AI readiness and inter-professional collaboration. *PLoS One.* 2025;20(5):e0322794. <https://doi.org/10.1371/journal.pone.0322794>. PMID: 40378129.

AUTHORSHIP CONTRIBUTIONS

Study design: Almeida YS, Cortez EA, Valente GSC.

Data collection: Almeida YS.

Data analysis: Almeida YS.

Data interpretation: Almeida YS.

All authors are responsible for the textual writing and critical review of the intellectual content, the final version published, and all ethical, legal, and scientific aspects related to the accuracy and integrity of the study.



Copyright © 2026 Online Brazilian Journal of Nursing

This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.