BRAIN. Broad Research in Artificial Intelligence and Neuroscience
e-ISSN: 2067-3957 | p-ISSN: 2068-0473
Covered in: Web of Science (ESCI); EBSCO; JERIH PLUS (hkdir.no); IndexCopernicus; Google Scholar; SHERPA/RoMEO; ArticleReach Direct; WorldCat; CrossRef; Peeref; Bridge of Knowledge (mostwiedzy.pl); abcdindex.com; Editage; Ingenta Connect Publication; OALib; scite.ai; Scholar9; Scientific and Technical Information Portal; FID Move; ADVANCED SCIENCES INDEX (European Science Evaluation Centre, neredataltics.org); ivySCI; exaly.com; Journal Selector Tool (letpub.com); Citefactor.org; fatcat!; ZDB catalogue; Catalogue SUDOC (abes.fr); OpenAlex; Wikidata; The ISSN Portal; Socolar; KVK-Volltitel (kit.edu)
2026, Volume 17, Issue 3, pages: 91-111
Submitted: May 24th, 2026 | *Accepted for publication:* August 23rd, 2026

The Post-ChatGPT Surge: Does AI-Health-Education Research Growth Reflect Evidence or Enthusiasm

Olasile Babatunde Adedoyin
University of Kyrenia, Kyrenia, Mersin 10, Turkey.
olasilebabatunde.adedoyin@kyrenia.edu.tr,
https://orcid.org/0000-0003-0166-2383

Fahriye Altinay
Societal Research and Development Centre, Institute of Graduate Studies, Faculty of Education, Near East University, Nicosia, North Cyprus, Mersin 10, Turkey.
fahriye.altinay@neu.edu.tr,
https://orcid.org/0000-0002-3861-6447

Gokmen Dagli
Faculty of Education, University of Kyrenia, Kyrenia, Mersin 10, Turkey.
gokmen.dagli@kyrenia.edu.tr,
https://orcid.org/0000-0002-4416-8310

Zehra Altinay
Societal Research and Development Centre, Faculty of Education, Near East University, Nicosia, Mersin 10, Turkey.
zehra.altinaygazi@neu.edu.tr,
https://orcid.org/0000-0002-6786-6787

** Corresponding author: Olasile Babatunde Adedoyin, University of Kyrenia, Kyrenia, Mersin 10, Turkey, olasilebabatunde.adedoyin@kyrenia.edu.tr*

Abstract: *Artificial intelligence (AI) research in health education has grown exponentially since the public release of ChatGPT in November 2022. However, questions remain regarding whether this growth represents a mature scholarly field or simply a self-feeding loop of publication growth. This bibliometric study examines the relationship between the increase in publications, citation impacts, and the thematic reorientation towards generative artificial intelligence after the launch of ChatGPT. Bibliographic records were extracted from Web of Science, Scopus, and PubMed and combined into one dataset comprising 4,358 publications after removing duplicates to address the limitation of the previous bibliometric analysis (based on only one bibliographic database). The R packages bibliometrix and biblioshiny were used to visualise and analyse the bibliometric data. Annual publications and citations, distribution of sources and authors, and thematic structure based on Author's Keywords (with keyword-field cleaning to eliminate indexing inconsistencies and combine related terms) were explored. The number of publications increased from 228 in 2022 to 1,550 in 2025, with an average annual growth of 56.48%. The growth peak was observed roughly one year after the release of ChatGPT, not right away, from 2023 to 2024. The average citation impact per document peaked in 2023, and it dropped steadily in all the following years, both for raw and time-normalised citation impact, with citation-impact concentration remaining locked onto the year immediately following ChatGPT's release. Thematic analysis revealed that in 2024, ChatGPT and other large language models started to dominate over machine learning, which previously accounted for the majority of publications. By employing factorial analysis, another cluster emerged from the documents, which associated particular Generative AI tools with measures of information quality (accuracy, readability). AI in health education research since ChatGPT's appearance may not only exhibit characteristics of either the evidence-generating or the enthusiastic type but also demonstrates tendencies suggesting that the former currently outweighs the latter. The results suggest a possible gap between the quantitative expansion of the literature and the consolidation of the field. The article explores the implications of these findings for research, funding, and health-education practice and outlines several productive directions for future bibliometric analysis.*

Keywords: *artificial intelligence; generative AI; ChatGPT; health education; patient education; bibliometric analysis; large language models.*

How to cite: Adedoyin, O. B., Altinay, F., Dagli, G., & Altinay, Z. (2026). The post-ChatGPT surge: Does AI-health-education research growth reflect evidence or enthusiasm? *BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 17*(3), 91-111. https://doi.org/10.70594/brain/17.3/6

©2026 Published by EduSoft Publishing. This is an open-access article under the CC BY-NC-ND license

1. Introduction

Artificial intelligence (AI) has evolved from a peripheral research topic in health professions to a topical issue with sustained research attention, and this shift did not happen gradually. Following a turning point triggered by OpenAI's public release of ChatGPT in November 2022, the volume of publications on generative AI (GenAI) in medical and health contexts increased to a degree few other technologies could match.

As documented by Temsah et al. (2023), the first nine months post-ChatGPT release increased PubMed-indexed records on ChatGPT-related literature to 1,000. A milestone that took Google fourteen years to reach in biomedical literature. The composition of that surge is what it fails to capture and what drives the current study.

Health education, as an interdisciplinary field that generally encompasses patient education, public and environmental health training, and medical and nursing curricula, has consequently become a prominent site for examining this expansion.

Currently, AI-enabled tools are utilised to stimulate clinical interactions, create patient-facing materials, customise individualised curricula for students, and respond to frequently asked patient questions across spheres, including orthopaedics and oncology (Guo et al., 2026).

The appeal makes sense because large language models (LLMs) offer scalability, personalisation, and constant availability, qualities that traditional health education has long struggled to provide due to clinician time constraints and patient variability in health literacy (Guo et al., 2026). However, appeal is not evidence, and the rate at which publications in this field are increasing raises a question that has so far received more conjecture than methodical investigation. Is the field's momentum being driven by an expanding base of rigorously designed, outcome-validated research, or by the sheer accessibility of writing about a technology that is easy to discuss and difficult to test?

A Scopus-based analysis of 3,404 documents published between 2013 and 2023 on AI in public health education by Ningrum and Muhtar (2024) reported an annual growth rate of 95.97% and identified federated learning, blockchain, and disease outbreak management as emerging thematic clusters.

Guo et al. (2026) analysed 428 Web of Science (WoS) indexed articles and reviews on AI in patient education, documenting that publication output increased from 40 papers in 2023 to 189 in 2024, with surgery-related applications dominating the field and GenAI keywords such as ChatGPT and LLMs anchoring the most recent thematic cluster. Similarly, Genc and Kocak (2025) studied 640 WoS-indexed documents published between 2020 and 2024 on AI in environmental and health education, finding a concentration of output in the most recent study year and identifying ChatGPT and machine learning (ML) among the most frequently occurring keywords.

The current study does not negate the methodological soundness of each of these studies within their scope. However, each also shared two similar limitations that directly influence this study's research question.

Primarily, all three studies relied on a bibliographic database, WoS in two cases and Scopus in the other. Because database coverage differs across disciplines and publication types, this design may omit clinically oriented literature that is more comprehensively represented in PubMed (Falagas et al., 2008).

In addition, the three studies did not code documents by evidentiary tier, and document-type exclusion criteria were not applied to exclude non-empirical content before analysis. Both Genc and Kocak (2025) and Guo et al. (2026) included the full range of article and review types indexed by WoS without distinguishing empirical output from commentary, perspective, or narrative-review content. Ningrum and Muhtar (2024) controlled their corpus to original articles. Thus, each study documented what the research domain is publishing about, but not what kind of publishing is actually taking place, or whether the growth curve each document is composed of is significantly lower-evidence content that a more targeted inclusion strategy would have excluded.

This gap goes beyond a methodological observation. Rather, it is at the center of an issue that has previously come up in other AI and health literature. For instance, Temsah et al. (2023) identified an early absence of empirical evidence even as enthusiasm for AI continued to build based on the authors' findings that indicated that one-third of the first 1,000 ChatGPT-related PubMed citations were editorials or commentaries rather than empirical studies. That finding influenced a deliberate methodological choice in this study. This concern was extended by Arzilli et al. (2025) into a broader discussion of scientific integrity, outlining the risk of an AI-powered 'infodemic' in which the ease of producing AI-assisted manuscripts, along with the current publish-or-perish incentives, could flood the literature with content that adds volume without adding validated knowledge.

The same issue was advanced from a design-standards perspective by Sallam et al. (2024), developing the METRICS checklist explicitly because the absence of evidence in consensus reporting guidelines for GenAI health studies was, in their assessment, already impeding the field's ability to synthesise evidence coherently.

Most pertinent to the current study, an umbrella review of ChatGPT-in-healthcare systematic reviews was conducted by Iqbal et al. (2025), and discovered that twelve of the seventeen included reviews had low quality ratings on the AMSTAR-2 instrument. This finding implies that the evidentiary issue is still present even in document types, such as systematic reviews, which are typically regarded as higher-tier evidence. When combined, these publications show that document-type filtering, even with careful application, is unlikely to completely distinguish between publication growth-driven output and evidence-generating research.

This study directly addresses these gaps by merging WoS, Scopus, and PubMed records, deduplicated into a single corpus, correcting for the single-database bias present in each of the three predecessor studies and restoring the clinically weighted coverage that WoS and Scopus alone tend to underrepresent.

This study spans January 2022 through 2026, a window that provides eleven months of pre-ChatGPT baseline data alongside the available post-release period through 2026, enabling comparison of publication behaviour, source concentration, authorship patterns, and thematic composition before and after the inflection point that Temsah et al. (2023) and others have identified as the field's defining discontinuity.

Accordingly, this study asks: even after excluding editorials and other non-empirical publication types, does the remaining growth in AI-health-education research correlate with indicators of methodological rigor, or does it continue to reflect a pattern more consistent with publication growth-driven scholarship, one dominated by scoping reviews, narrative commentary framed as original research, and thematic clustering around promissory rather than validated applications of GenAI.

In pursuing this question, the study aims to move the literature beyond descriptive mapping, toward a diagnostic account of what the post-ChatGPT surge in AI-health-education scholarship actually consists of, even under a more conservative inclusion strategy than any prior study in this space has applied.

The following research questions were raised to achieve the research aims;

1. What is the production and impact trajectory of studies published on AI in health education between 2022 and 2026, and how does this trajectory differ before and after ChatGPT's November 2022 release?
2. What are the dominant and emerging themes in AI-health-education research, and has the field's thematic composition shifted from methodologically grounded concepts toward GenAI-specific and promissory language following ChatGPT's November 2022 release?

2. Methodology

2.1. Research Design

This study employed a bibliometric research design to examine publication and citation patterns in AI and health education research published between January 2022 and August 2026. This research design was adopted because it enables systematic, quantitative examination of large scholarly corpora across multiple dimensions, such as publication volume, citation impact, source concentration, authorship structure, geographic distribution, and thematic composition, at a scale that narrative review cannot achieve (Donthu et al., 2021).

Citation impact reflects how frequently a publication, or a set of publications from a specific year, is cited by other scholars. This is usually applied as a proxy for the work's influence or prestige within the field. In this study, both raw citation counts and citations per year are reported as citation impact, since the latter better reflects differences in citation practices for older versus more recent works.

In accordance with the fundamental research question of the current study, a sequential comparison across the pre- and post-ChatGPT release periods to determine whether growth, concentration, and thematic patterns shifted following the technology's public release in November 2022 was incorporated into the study design.

2.2. Data Sources and Search Strategy

Scopus, WoS, and PubMed served as the three databases for bibliographic records and were jointly queried, compared with prior bibliometric studies that use individual queries. Adopting a single citation index has been shown to limit results to journal coverage and the database's discipline. Scopus and WoS are substantially different in coverage with respect to the research domain. This variation has been shown to have a significant effect on bibliometric results depending on the database or combination adopted (Caputo & Kargina, 2022).

In contrast, PubMed offers more all-inclusive indexing of clinical and biomedical journals than WoS or Scopus, especially for empirical research centred on patient outcomes (Falagas et al., 2008). To reduce the risk of disciplinary bias associated with a single-database search, this study combined all three databases.

The search strategy incorporated combined keywords related to AI technologies and health education. The data were subjected to inclusion-exclusion criteria during data collection, which excluded letters, books, book series, erratum, research support, extramural, and trade publications, thereby limiting eligible sources to articles, reviews, and conference/proceedings papers.

The exclusion of editorials was justified by the findings of Temsah et al. (2023) that one-third of the first 1,000 PubMed ChatGPT-related citations were editorials/commentaries. The authors argue that editorials/commentaries constitute a significant proportion of health literature output immediately following the introduction of a disruptive technology. Therefore, instead of developing a strategy to include or exclude them in the analysis, the option of removing such entries from the initial database was chosen.

The data collection procedure followed the recommendations of the PRISMA 2020 statement (Page et al., 2021) to report screening and selection processes in accordance with the inclusion-exclusion criteria, while mentioning the search keywords used for database screening. It should be noted that PRISMA does not apply to bibliometric field mapping studies. Thus, the current analysis followed the guidelines for reporting systematic reviews, focusing on the elements relevant to the search process.

Table 1 below contains the data collection search strings. Searches were restricted to records indexed between January 1, 2022, and August 12, 2026, and English-language publications only. All three database searches were conducted on August 12, 2026, to preserve time-based consistency across sources.

Table 1. Data Collection Search Strings

Databases	Search Strings
Scopus	(TITLE-ABS-KEY(("health education" OR "health promotion" OR "health literacy" OR "patient education" OR "hygiene education" OR "wellness education" OR "community health education")) AND TITLE-ABS-KEY(("artificial intelligence" OR "machine learning" OR "deep learning" OR "generative ai" OR "large language model*" OR "chatgpt" OR "nlp" OR "llm"))) AND NOT TITLE-ABS-KEY(("medical education" OR "nursing education" OR "medical school*" OR "clinical training" OR "residency training"))
WoS	TS=("health education" OR "health promotion" OR "health literacy" OR "patient education" OR "hygiene education" OR "wellness education" OR "community health education") AND TS=("artificial intelligence" OR "machine learning" OR "deep learning" OR "generative AI" OR "large language model*" OR "ChatGPT" OR "natural language processing" OR "NLP" OR "LLM") NOT TS=("medical education" OR "nursing education" OR "medical school*" OR "clinical training" OR "residency training")
PubMed	(("Health Education"[Mesh] OR "Health Promotion"[Mesh] OR "Health Literacy"[Mesh] OR "Patient Education as Topic"[Mesh] OR "health education"[tiab] OR "health promotion"[tiab] OR "health literacy"[tiab] OR "patient education"[tiab] OR "hygiene education"[tiab] OR "wellness education"[tiab] OR "community health education"[tiab]) AND ("Artificial Intelligence"[Mesh] OR "Machine Learning"[Mesh] OR "Deep Learning"[Mesh] OR "Natural Language Processing"[Mesh] OR "artificial intelligence"[tiab] OR "machine learning"[tiab] OR "deep learning"[tiab] OR "generative ai"[tiab] OR "large language model*"[tiab] OR "chatgpt"[tiab] OR "nlp"[tiab] OR "llm"[tiab])) NOT ("Education, Medical"[Mesh] OR "Education, Nursing"[Mesh] OR "medical education"[tiab] OR "nursing education"[tiab] OR "medical school*"[tiab] OR "clinical training"[tiab] OR "residency training"[tiab])

2.3. Data Merging and Deduplication

The records retrieved from the WoS, Scopus, and PubMed databases were uploaded into R software using the bibliometrix package (Aria & Cuccurullo, 2017). The analysis was conducted using the Shiny-based interface of the bibliometrix package called 'biblioshiny'. This application provides an integrated view of the bibliometric analysis results. The records were deduplicated using a DOI-matching algorithm. The choice of this particular algorithm reflects the current best practice in the field: merging multiple bibliometric databases typically involves using DOIs for merging and relying on secondary matching criteria (Nikolić et al., 2024). However, there was a secondary deduplication step, which relied on matching the title, the first author's surname, and the year of publication.

This was done because there are cases where a record lacks a DOI, and this is particularly common for the PubMed database, which contains preprints and conference abstracts. While DOI-based matching is generally the most reliable option, it is important to note that this database has a significant number of entries that lack DOIs, and the reliability of the match is lower for non-DOI entries. This consideration applies to other databases, but the problem is most acute for PubMed, which is why it is noted. The initial combined corpus contained 7,933 entries (WoS = 2,742, Scopus = 4,677, PubMed = 514). After deduplication, the merged corpus contained 4,358 entries, of which 45.06% were duplicates.

It is also important to note that one limitation of the current approach, as detailed in the merged corpus description, concerns the fact that PubMed does not contain citation data. Thus, any analysis that relies on citations, including the most cited documents per year, annual citation rate, source local impact, and author local impact, is based solely on the WoS and Scopus records. The ones found in PubMed only contribute to the overall number of publications.

2.4. Analytical Software

All bibliometric analyses were performed using the bibliometrix R package through 'biblioshiny' (Aria & Cuccurullo, 2017), with references to the latest version of the package's science-mapping workflow (Aria et al., 2026). The bibliometrix tool has been used in preference to other available software, such as VOSviewer and CiteSpace, for its ability to perform bibliometric

analyses on combined databases and report production, impact, and thematic indicators in an integrated analytical framework.

2.5. Analytic Procedures

This study involved two stages, namely production and impact trajectory and topic and semantic evolution. In the first stage, the number of documents published per year and their citation impact were determined for each year for the entire period (2022-2026). The former period, particularly the year 2023, is regarded as ChatGPT's launch period, with a significant implication for the research field. The study includes a special note that the values for 2026 are preliminary since this year has been only partially covered by the search. Thus, for the last period, the overall number of articles rather than the average one was taken.

For the majority of the documents retrieved worldwide, the year of publication was analysed to identify if the concentration of highly cited works was determined before ChatGPT's appearance or if the recent studies started to gather more attention despite a shorter publication period.

The second stage included the identification of the most frequent words, word cloud, and tree map visualisation. The word frequency distribution helped establish if there were any trends over time, which terms gained or lost ground, and when GenAI-related terminology began to dominate the corpus. This information was critical to understand if the field used a specific set of keywords related to AI before ChatGPT's introduction and how the dominance of these terms changed over time.

Besides frequency, the presence of GenAI-related terms in the co-word network was evaluated to identify if they formed either a central concept or a peripheral category within the field. The thematic map visualised the field's concepts based on their density and centrality to reveal motor, basic, niche, and declining areas using the default settings of bibliometrix's strategic-diagram construction.

Thematic density reflects how much a theme is internally integrated and cohesive based on the extent to which keywords in a theme co-occur with each other. Centrality, on the other hand, reflects how much a keyword or a theme is integrated with other themes in the field. For the former, a theme with a high-density value is well-established: it has a stable internal structure and a set of keywords that are specific to it. A theme with a low-density value is emerging: the set of keywords associated with it is fluid and has not yet stabilised.

For the latter, a theme that has a high centrality value is closely related to many other themes in the field and represents a cross-cutting concern in the literature. In contrast, a theme with a low centrality value is only distantly related to other themes and is discussed in isolation. Factorial analysis was applied to the keyword co-occurrence matrix using multiple correspondence analysis (MCA) to identify if the field's concepts shifted into several groups and if the group sizes changed over time.

2.6. Temporal Comparison

The corpus spans January 2022 (n = 269 records) through 2026, and ChatGPT was made public in November 2022; the year 2022 was taken as the pre-release baseline, while the following four years were considered post-release for this analysis. ChatGPT was released on November 30, 2022, meaning the 2022 calendar year is composed almost entirely (eleven of twelve months) of pre-release publications, and is therefore treated as a single baseline year rather than split into pre/post sub-periods. The analyses described above were carried out for the entire corpus, and where possible, for its baseline and post-release subperiods, to assess the differences between these two sets beyond eyeballing the trends yearly.

2.7. Ethical Considerations

This study only involved analysing publicly available bibliographic information and citation data, and did not require any human subjects testing or involvement with confidential material, no ethical approval was required.

2.8. Limitations of the Analytic Approach

Notably, there are two potential limitations to the current approach. First, while editorial commentaries, letters, notes, books, book series, book reviews, conference reviews, and trade journal documents were excluded at the database level, this study cannot distinguish between empirical and narrative literature types beyond conventional categories such as reviews, because narrative reviews, scoping reviews, and perspectives published as regular articles are not differentiated by any of the three databases used. Second, the current analyses that rely on citation data are limited by what is included in the WoS and Scopus databases, since PubMed does not provide citation data, and both of these resources only include citations up to their most recent update, which usually has a delay of about a year after the actual date of publication for any given paper. Bibliometric indicators (citation trajectory, thematic density, keyword co-occurrence) are used as proxies for evidentiary composition, not as direct measures of study quality, since our design does not include manual assessment of individual studies.

3. Findings

3.1. Production and Impact Trajectory from 2022 to 2026

Table 2 captures the annual production distribution of studies published on AI in health education by year, and Table 3 captures annual citation performance, indicating that the studied corpus accounts for 4,358 documents in total during the period under investigation. The growth rate of this corpus is automatically calculated as 56.48% per annum, whereas the average number of citations per document is 10.37.

3.1.1. Production Trajectory

Firstly, it should be noted that in the context of the study, the production distribution of studies demonstrates a consistent exponential growth trend. In particular, there were 228, 348, 851, and 1,550 studies published in 2022, 2023, 2024, and 2025, respectively, which demonstrates a significant increase of 52.6%, 144.5%, and 82.1% from one year to another. While the number of articles published in 2026 is 1,367, meaning that, compared to 2023, it already covers a full year of research, whereas, compared to 2024, it is only slightly lower, which suggests that the trend is unlikely to slow down significantly ahead of 2026, which might allow to conclude that the growth rate has not diminished at the time when the data was collected for this study. To put the growth rate in perspective, Ningrum and Muhtar (2024) only managed to achieve 95.97% for the narrower scope of AI-in-public-health-education during 2013-2023, whereas the tendency observed for the broader topic of AI-in-health education is somewhat lower, yet still very considerable.

Table 2. Annual Production Distribution

Year	Articles
2022	228
2023	348
2024	851
2025	1550
2026	1367

As for the study's central research question, even within the limitations of the 2022-2026 span, it is arguable that the growth pattern does not appear to have been significantly affected by either ChatGPT’s launch at the end of 2022 or the onset of the pandemic in 2020. In particular, while the increase of 52.6% across 2022-2023 is actually the lowest recorded since 2023, the jump of 144.5% from 2023 to 2024 appears to be much higher, meaning that a significant portion of the

research community needed almost a year to adjust to the new developments in the area to produce new scholarly works, which is why the growth spurt came later.

3.1.2. Citation and Impact Trajectory

This is probably the most telling difference from the production curve, and the evidence-generating research versus publication-growth dichotomy is clearly taking shape. The mean total number of citations per article peaks in 2023 (42.41) before sharply declining, with no sign of recovery, to 18.47 for 2024, 5.35 for 2025, and 0.68 for 2026. For the most part, this is because recent years have less time to accrue citations (five citable years for 2022 and only one for 2026). Still, even when adjusting for this bias using the annual mean, the peak in 2023 stands out, with numbers of 4.7, 10.6, 6.16, 2.67, and 0.68 for 2022, 2023, 2024, 2025, and 2026, respectively. If the sole reason for variance between the cohorts was the number of years since publication, the annual mean would be close to, or slightly above, zero in all cases. Therefore, using citation numbers as a metric for impact, it is reasonable to conclude that 2023 was indeed a peak year for the field, after which impact per article has declined.

Table 3. Annual Citation Performance

Year	Mean TC per Article	N	Mean TC per Year	Citable Years
2022	23.5	228	4.7	5
2023	42.41	348	10.6	4
2024	18.47	851	6.16	3
2025	5.35	1550	2.67	2
2026	0.68	1367	0.68	1

In short, 2023 was the year when the field peaked in terms of both volume and impact, with impact per article driving the opposite trajectory to the overall production volume.

3.1.2. Most-cited Studies

As shown in Table 4, which captures the top 19 most-cited articles, the trend is consistent with the peak year hypothesis; indeed, twelve out of nineteen most-cited studies were published in 2023, three in 2022, and four in 2024. Notably, there are no most-cited documents for 2025 and 2026, as the fields require some time to build up citation presence. Thus, while overall production and impact per article peak in 2023, the distribution of the most-cited documents is clearly skewed toward the year of the paradigm's introduction, supporting the impact concentration hypothesis.

Table 4. Most-Cited Articles

Articles	DOI	Total Citations	TC per Year	Normalized TC
ALOWAIS S, 2023, 1	10.1186/s12909-023-04698-z	1809	452.3	42.7
SALLAM M, 2023, 6	10.3390/healthcare11060887	1731	432.8	40.8
SANTODOMINGO-RUBIDO J, 2022, CONTACT LENS ANTERIOR EYE	10.1016/j.clae.2021.101559	643	128.6	27.4
YEO Y, 2023, 3	10.3350/cmh.2023.0089	515	128.8	12.1
LI H, 2023, NPJ DIGIT MED	10.1038/s41746-023-00979-5	445	111.3	10.5
AGGARWAL A, 2023, J MED INTERNET RES	10.2196/40789	444	111.0	10.5
SORBARA M, 2022, NAT REV MICROBIOL	10.1038/s41579-021-00667-9	396	79.2	16.9
KHOGALI H, 2023, TECHNOL SOC	10.1016/j.techsoc.2023.102232	301	75.3	7.1

BHAYANA R, 2024, RADIOLOGY	10.1148/radiol.232756	285	95.0	15.4
GUAN Z, 2023, CELL REP MED	10.1016/j.xcrm.2023.101213	268	67.0	6.3
SUJITH A, 2022, NEUROSCI INFORM	10.1016/j.neuri.2021.100028	264	52.8	11.2
LYU Q, 2023, VIS COMPUT IND BIOME	10.1186/s42492-023-00136-5	250	62.5	5.9
VAN D K, 2024, LANCET PUBLIC HEALTH	10.1016/S2468-2667(24)00055-0	245	81.7	13.3
DAMAŠEVIČIUS R, 2023, INFORMATION	10.3390/info14020105	243	60.75	5.7
VARGHESE C, 2024, NAT MED	10.1038/s41591-024-02970-3	237	79	12.8
SAMAAN J, 2023, OBES SURG	10.1007/s11695-023-06603-5	234	58.5	5.5
PAN A, 2023, 10	10.1001/jamaoncol.2023.2947	222	55.5	5.2
SHAHSAVAR Y, 2023, JMIR HUM FACTORS	10.2196/47564	218	54.5	5.1
ESMAEILZADEH P, 2024, ARTIF INTELL MED	10.1016/j.artmed.2024.102861	212	70.7	11.5

The most-cited article among all is also a 2023 publication by Alowais et al. (2023). This study is also the most-cited in the bibliometric analysis of AI and patient education by Guo et al. (2026), and, judging by the classification of the study as a review, this suggests that the field's most impactful contribution in terms of volume of impact appears to be dominated by non-empirical studies.

3.2. Thematic Composition and Evolution of AI in Health Education from 2022 to 2026

Thematic analysis was conducted using the Author's Keywords (DE) field after merging synonymous keywords (large language model/models/models, llm/llms, AI/artificial intelligence). As a result, 50 keywords reached the required threshold for network- and map-based analysis.

3.2.1. Dominant Themes

Across the combined corpus, the most frequent keyword was "artificial intelligence" accumulating 1,888 appearances with 25% of total keyword volume followed by "large language model" appearing 682 with 9%, "patient education" amassing 663 occurrences with 9%, "chatgpt", and "machine learning" appearing 627, and 390, with 8% and 5% of total keyword volume, respectively (See Figure 1).

The following set of keywords included "health literacy" raking in 316 occurrences with 4%, "readability" pulling in 232 appearances with 3%, "digital health" netting 195 occurrences with 3%, "generative artificial intelligence" earning 170 appearances with 2%, "chatbot" amassing 159 occurrences with 2%, and "natural language processing" claiming 155 appearances with 2%. Specifically, the terms related to GenAI, including LLM, ChatGPT, generative artificial intelligence, chatbot, and Gemini and DeepSeek as the lower-frequency keywords, accounted for approximately 25% of total keywords volume, comparable to that of the term "artificial intelligence".

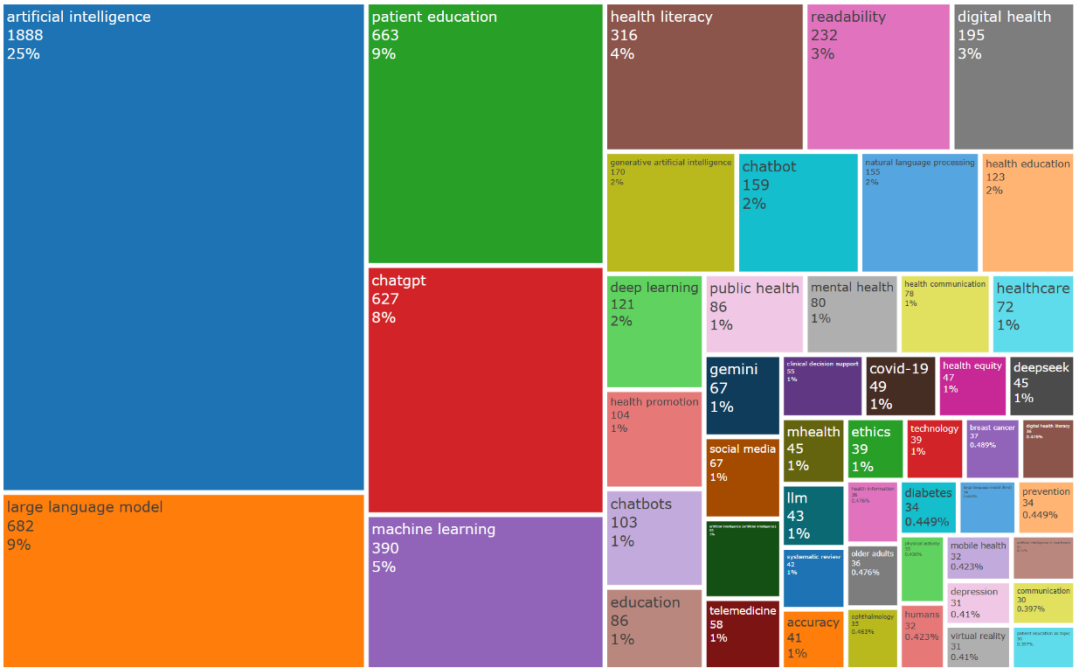


Figure 1. Dominant Themes

3.2.2. Temporal Evolution of Keyword Frequency

Annual (non-cumulative) keyword frequencies, extracted from the cumulative Words' Frequency over Time visualisation, are provided in Table 6 below. The term "machine learning", once one of the most widespread AI-related terms, appeared 36, 45, 85, 131, and 88 times in 2022, 2023, 2024, 2025, and (up to August) in 2026, respectively. The term "ChatGPT" was absent in 2022, but appeared 35, 189, 249, and 150 times in 2023, 2024, 2025, and (up to August) in 2026, respectively. Meanwhile, "large language model" was first detected in 2023, with 11, 100, 262, and 309 occurrences in 2023, 2024, 2025, and (up to August) in 2026, respectively.

The number of articles including the term "ChatGPT" started overtaking the number of articles that included the term "machine learning" in 2024 (189 vs. 85) and remained higher since then. Similarly, the number of articles that include the term "large language model" became higher than the number of articles that included "machine learning" in 2024 (100 vs. 85) and continued increasing at a faster pace than the usage of the older term, ultimately reaching 309 in 2026 (vs. 150 for "ChatGPT").

Table 6. Words' Frequency over Time

Keywords	**Years**				
	2022	**2023**	**2024**	**2025**	**2026**
Artificial Intelligence	37	164	545	1224	1880
Large Language Model	0	11	111	373	682
Patient Education	6	32	151	402	658
Chatgpt	0	35	224	473	623
Machine Learning	36	81	166	297	385
Health Literacy	17	35	91	200	313
Readability	0	6	52	129	231
Digital Health	10	26	49	112	195
Generative Artificial Intelligence	0	2	20	77	168
Chatbot	3	25	63	125	158

3.2.3. Trending Topics

Through trending topic analysis that examines terms according to the chronological distribution of their occurrences rather than their overall frequency, thirteen keywords that met the selection criteria were identified. Keywords with the earliest median publication years were "forecasting" (2022), "COVID-19", "big data", and "behaviour change" (2023). The keywords with the latest median years were "artificial intelligence", "large language model", and "patient education" (2025), and "generative artificial intelligence", "clinical decision support", and "DeepSeek" (2026), whereas the years for "mhealth", "technology", and "conversational agent" were 2024. "ChatGPT" was excluded from this analysis, as it did not meet the selection criteria, thus limiting its inclusion to a few studies throughout the timeline of the research.

3.2.4. Co-Word Network

The results of the keyword co-occurrence analysis identified two main clusters (See Figure 2). The first cluster included "machine learning" (betweenness centrality = 34.19), "digital health" (20.26), "health promotion", "public health", "mental health", "social media", "telemedicine", "COVID-19", "mhealth", "systematic review", "ethics", "technology", and several other related demographic and condition-specific keywords. The second cluster comprised "artificial intelligence" (betweenness centrality = 438.28), "large language model" (62.72), "patient education" (42.36), "ChatGPT" (40.59), "health literacy" (12.71), "readability", "generative artificial intelligence", "chatbot", "natural language processing", and "health education". Betweenness centrality, PageRank (0.186), and closeness (0.0204) were highest for "artificial intelligence", suggesting its significant role in bridging other terms.

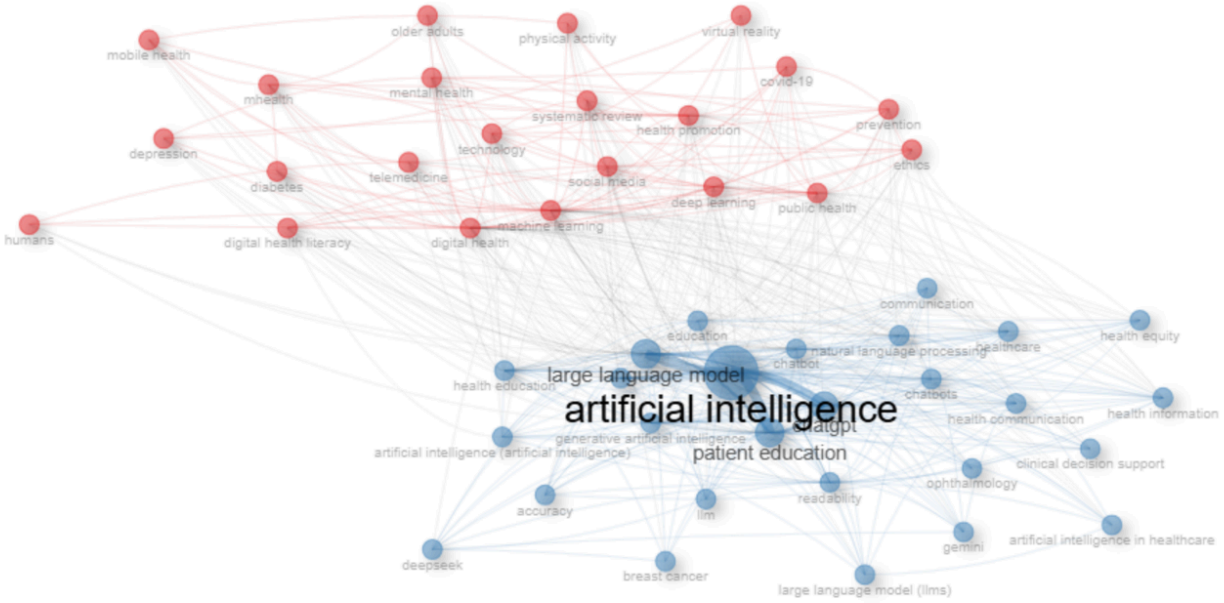


Figure 2. Co-word Network

3.2.5. Thematic Map

The strategic diagram (density-centrality mapping) also revealed three distinct keyword clusters that occupied different quadrants (See Figure 3). The first cluster contained no keywords that fell under the Motor Themes quadrant (high density and high centrality). The group 'Artificial intelligence/machine learning/health literacy' (number of occurrences: artificial intelligence = 1,774, machine learning = 389, health literacy = 314, with these counts representing only the keywords included in the particular cluster) was located in the Basic Themes quadrant, suggesting its strong presence in the analysed field of study while not possessing high thematic development. The second group 'Digital health/health promotion/public health' was located near the Niche/Motor

quadrant. ‘Large language model/patient education/chatgpt’ (number of occurrences: ‘Large language model/patient education/chatgpt’ = 678, ‘Large language model/ patient education/chatgpt’ = 660, ‘Large language model/patient education/chatgpt’ = 624, with these counts representing only the keywords included in the particular cluster) represented the last cluster located in the Emerging or Declining Themes quadrant, which was the lowest ranked according to both density and centrality.

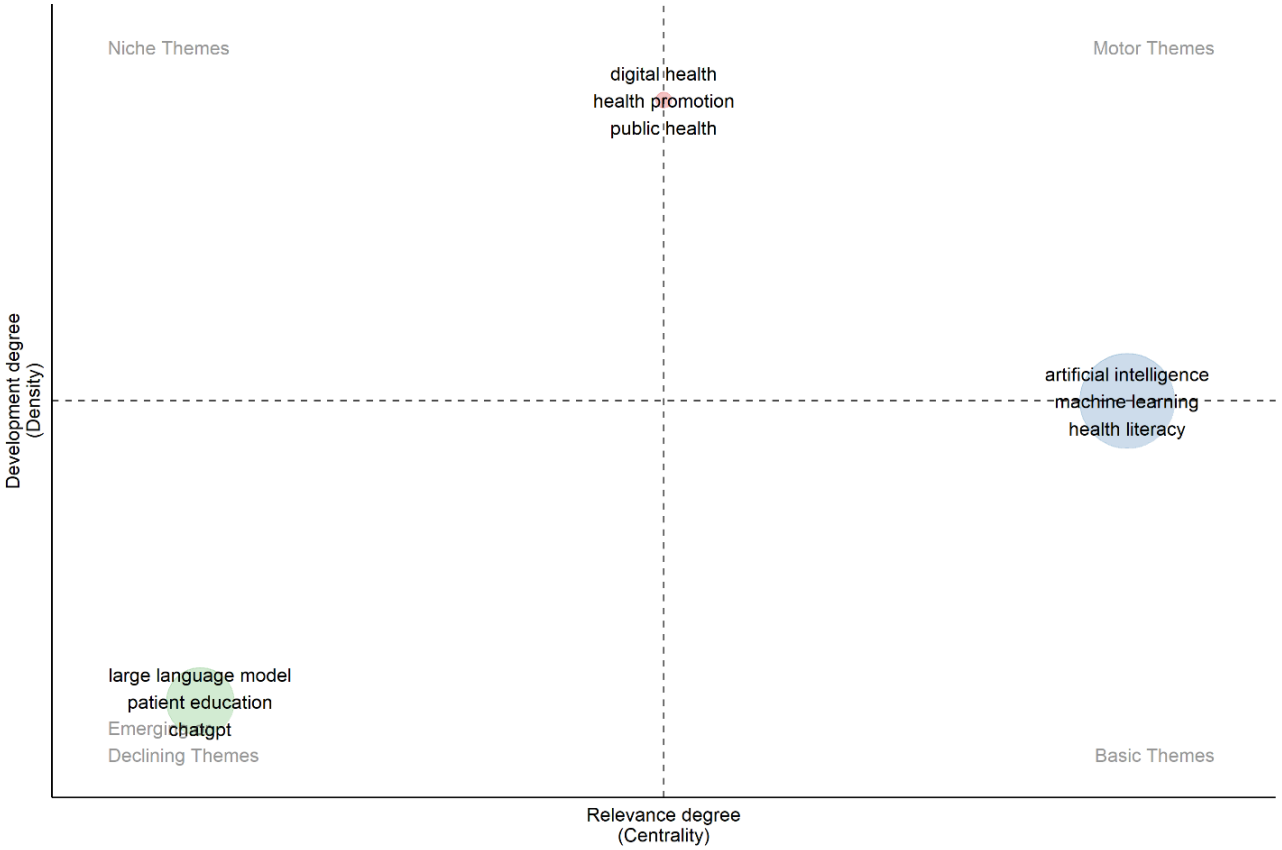


Figure 3. Thematic Map

3.2.6. Factorial Analysis

The MCA of the 50 most frequent keywords with a cluster number set to find multiple groups revealed four conceptual clusters (Figure 4). Dimension 1 was responsible for 40.08% of the variance, while Dimension 2 accounted for 9.76%.

Cluster 1, which was the largest one in terms of the number of keywords, included “artificial intelligence”, “machine learning”, “health literacy”, “digital health”, “generative artificial intelligence”, “health education”, “deep learning”, “health promotion”, “public health”, “mental health”, “healthcare”, “telemedicine”, “COVID-19”, “ethics”, and most of the demographic and condition-related keywords. Cluster 2 comprised “large language model”, “patient education”, “ChatGPT”, “readability”, “Gemini”, “DeepSeek”, “accuracy”, “ophthalmology”, “gpt-4”, “google”, “patient information”, “google gemini”, and “patient education materials”, a grouping combining specific GenAI product and model names with terms denoting information-quality evaluation (“readability”, “accuracy”).

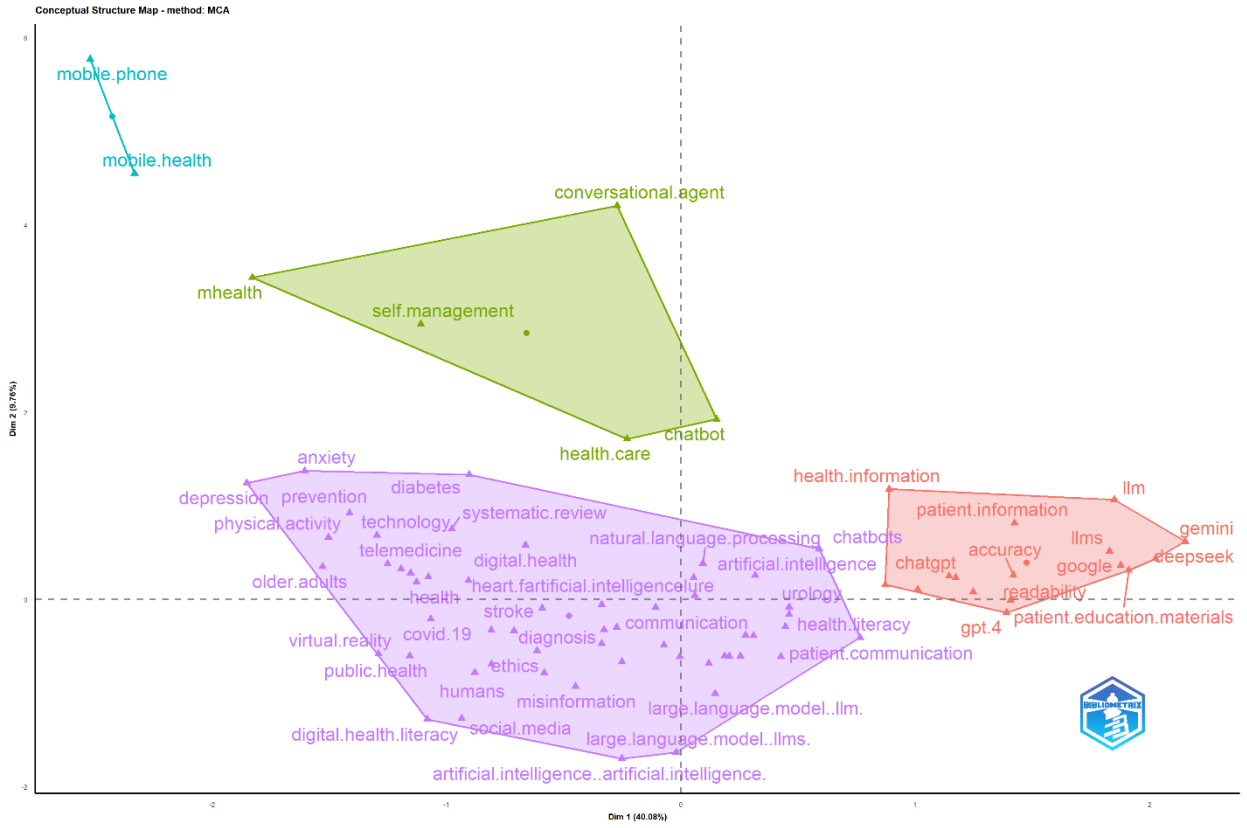


Figure 4. Conceptual Structure Map

Cluster 3 comprised “chatbot”, “mhealth”, “self-management”, “conversational agent”, and “health care”. Cluster 4 comprised “mobile health” and “mobile phone”, positioned as clear outliers on Dimension 2 (4.55 and 5.77, respectively) relative to all other keywords in the analysis.

4. Discussion

4.1. Production and Impact Trajectory from 2022 to 2026

Citation-accumulation lag is a known phenomenon common to any emerging field, and it is important to recognise where this study’s limitations begin and end. In Paleti et al. (2025) bibliometric study of the Asian orthopaedic and sports medicine studies between 1996 and 2022, a similar bibliometric trend was observed for their subject area, with the most recent years having significantly dropped in terms of citation counts, and they attribute this to the same effect of citation-accumulation lag and not a decline in research quality. This study is partially similar to Paleti et al. (2025) in this regard, as a part of the drop seen in 2024, 2025, and particularly noticeable in the partial 2026 cohort, can be attributed to the same causes.

However, the difference between this curve and the one seen in Paleti et al.'s (2025) research is evident when comparing the two cohorts with a sufficient amount of time between their publication dates, namely, 2022 and 2023. As of the date of this study, both had been out for over a year, and thus, should have been subject to similar levels of citation-accumulation lag. However, while the 2023 output’s average citation count per article is roughly double that of the 2022 output, the following cohorts exhibit a significantly lower count. A lag effect alone would explain a steady decrease between the 2023 and 2024, 2025, and 2026 cohorts, but not the significant peak in the middle of the window attributable to the ChatGPT launch date.

This pattern is broadly consistent with concerns raised elsewhere in the literature on AI and health education scholarship since ChatGPT’s release. Temsah et al. (2023) documented that ChatGPT-related PubMed citations reached 1,000 within nine months of the tool’s release. A pace far exceeding prior health-technology adoption curves, and they found that one-third of those early

citations were editorials or commentaries rather than empirical research, which are swiftly published outputs that would be expected to generate an early citation burst without corresponding scientific depth. Arzilli et al. (2025) extended this into a broader concern about an AI-driven publishing infodemic, in which publish-or-perish incentives combined with the ease of AI-assisted manuscript production risk flooding the literature with volume that does not track validated knowledge. Sallam et al. (2024) approached the same problem from a design-standards angle, developing the METRICS checklist specifically because the absence of consensus reporting guidelines for GenAI health studies was already impeding coherent evidence synthesis as early as 2024. Iqbal et al. (2025), in an umbrella review of ChatGPT-in-healthcare systematic reviews, found twelve of seventeen included reviews rated low quality on AMSTAR-2, suggesting the evidentiary concern extends into the secondary literature meant to synthesise primary findings, not just the primary literature itself.

Taken together, these articles provide context and support for the observed citation trends in this study. Within this body of evidence, this study's findings align with the trends observed by Arzilli et al. (2025) and Temsah et al. (2023): the field saw a significant increase in research volume right after ChatGPT's release, much of which consisted of reviews, and thus produced a significant citation peak right after the peak in research output, after which the field's literature failed to exhibit the same level of impact per published article.

Cheng et al. 's (2024) umbrella review of bibliometric studies of the growth of the life sciences literature acknowledges the need for international guidelines for managing the growth of the scientific literature, particularly in certain areas of medicine. The need for such guidelines has become evident, and the present study's findings add to the evidence, with the field of health education being a pressing concern within the broader field of health scholarship. Notably, the production growth curve peak observed by Cheng et al. (2024) appears to have occurred around the same time this study's citation curve did, around the early months of 2023: the former peaked at "+144.5% increase in publication volume per year" (Cheng et al., 2024).

The World Health Organization (2024) guidance issued at the beginning of 2024 on large multi-modal models in health, which responded directly to the rise of GenAI, such as ChatGPT, in healthcare, appears to coincide with the peak observed in this study. Compared to the bibliometric study by Ningrum and Muhtar (2024) on AI in public health education, this study's 56.48% growth rate is significantly lower than the 95.97% growth rate observed between 2013 and 2023 in the latter article's field, but closer to Guo et al. (2026)'s 40 to 189 publications increase between 2023 and 2024 in the AI-assisted patient education field. Genc and Kocak (2025)'s analysis of studies on AI applications in education, environment, health, and other domains also observes a peak in a single year, 2024, across all four fields. Notably, none of these articles explore the relationship between the growth in literature production and citation impact, only observing the two as separate trends.

In comparison, this study observes and discusses the evidence of a volume-impact disconnect in the literature on AI in health education. Thus, its findings complement the prior three articles by providing additional context and analysis. This study's discussion of the impact of the divergence on policy, practice, and research is an important addition to the findings in Ningrum and Muhtar (2024), Guo et al. (2026), and Genc and Kocak (2025).

4.2. Thematic Composition and Evolution of AI in Health Education from 2022 to 2026

The findings answer the second research question of the study more thoroughly than expected by providing several key themes. The themes that are dominant in the corpus, in descending order, are artificial intelligence, large language model, patient education, ChatGPT, and machine learning, with health literacy, readability, digital health, generative artificial intelligence, and chatbot forming the next tier. The emerging themes that appeared in the corpus at certain points in time, rather than frequently throughout, are generative artificial intelligence, clinical decision

support, and DeepSeek, with the last one being noteworthy as the first sign that the field is broadening beyond one particular product (ChatGPT) to consider the whole space of GenAI.

On the question of whether thematic composition has shifted toward GenAI-specific vocabulary, the evidence suggests that it has, and that the shift is both substantial and dateable. Annual mentions of ChatGPT and LLM each crossed the threshold of machine learning, which is the field's previous dominant term, emerging in 2024, a full year after ChatGPT's launch in November 2022. And this shift is not spurious; it suggests a genuine shift in thematic composition, rather than simply a change in the size of the corpus as a whole, because it appears in the pattern of annual (non-cumulative) mentions.

On the second part of the research question that concentrates on whether this shift is towards the promissory rather than the methodologically grounded language, the findings are more nuanced than the research question suggests. There are two findings that together suggest that the answer can be yes, but only partially. First, while ChatGPT, Gemini, DeepSeek, and GPT-4 were clustered together with terms related to readability, accuracy, and other information quality metrics, the words related to speculation or promotion were not found. This suggests that there may be a sub-literature within the GenAI-specific literature that actually evaluates the outputs of these tools according to certain criteria. Second, the cluster containing LLM, patient education, and ChatGPT was found in the Emerging or Declining Themes quadrant, the area of the map that represents the lowest possible centrality and density. Placing this cluster in this quadrant means that, while it may be growing and widespread, it may not be central to the discussion within the field. Had GenAI-specific literature completely taken over the discourse, the cluster may have been in the Motor Themes quadrant, the most central and dense one on the map. The most supported conclusion is that the field's vocabulary may have shifted towards GenAI-specific terminology in terms of both volume and recency. However, this shift does not involve a corresponding change in the thematic centrality and density of the discourse.

This finding echoes similar, and in part supportive, analyses in a small but growing literature applying discourse and topic-modeling approaches to academic writing about ChatGPT and related GenAI, also published in the window that this current study covers. In particular, Tao and Shen's (2025) analysis of 1,227 SSCI-indexed abstracts discussing ChatGPT in the social sciences during the same period (November 2022-November 2024) using topic modeling and sentiment analysis techniques also finds differing patterns of term use across six identified thematic categories, including AI and technology communication and education, suggesting that scholarly discussion of ChatGPT is not uniformly positive or promotional across themes. Similarly, an analysis in a comparable domain finding differentiated use of terms associated with promotional language provides further support that the factorial analysis presented here also finds a genuine underlying evaluative dimension in the health-education corpus, not merely an artifact of the keyword-cleaning procedure.

In a narrative review that focused on the impact of ChatGPT and related GenAI tools on public health research and scientific communication by Yoga Ratnam (2025), the scholar noted that while the field's beginnings were undeniably linked with the appearance of ChatGPT in November 2022, the subsequent major upgrades, including ChatGPT-4 in March 2023 and ChatGPT-4o in May 2024, contributed to a further leap in its adoption and capabilities, reflected in the literature. This is an important nuance that should be added to the present analysis, as the current finding of the ChatGPT and LLMs' annual mentions peaking in 2024 rather than right after the initial appearance of ChatGPT may be explained precisely by the emergence of the aforementioned upgrades. In fact, observations by Yoga Ratnam (2025) point to the possibility of GPT-4 and GPT-4o's release being a more significant inflection point, possibly fuelling the rise of the field's interest and volume of research. Thus, further investigations would benefit from bearing this possibility in mind and distinguishing between different versions of ChatGPT when undertaking similar analyses.

In a scoping review of the benefits, challenges, and uses of GenAI in healthcare by Moulaei et al. (2024), these researchers documented findings largely similar to the thematic map of this

study. Indeed, the authors' broad but shallow findings correspond closely to the current themes identified in the health education literature, and it is possible that the similar trends detected across the healthcare-related literature, as opposed to being specific to the domain of health education, are in fact characteristic of the general GenAI landscape. This possibility goes beyond the scope of this particular study and is in turn linked to further research opportunities.

The predominance of evaluative terms detected in this study, particularly readability and accuracy, around specific named GenAI tools, stands to reason in the context of Sallam et al. (2024) discussing the lack of standardised reporting in GenAI health literature, and Iqbal et al. (2025) noting that the majority of ChatGPT-related systematic reviews in healthcare were rated as low quality according to AMSTAR-2. Together, these findings highlight the presence of a certain degree of evaluative thrust in the literature, although its expression is yet to be systematic or standardised across studies, thus resulting in a low volume of high-quality systematic reviews, precisely the gap a thematic map's low-density reading would predict.

5. Conclusion

This study aimed to answer two related questions about AI in health-education research in the four years after the launch of ChatGPT: what form has its development and impact taken, and what have they developed into. Collectively, the findings demonstrate one overall conclusion that neither confirms nor refutes the concern posed in the title of the study.

In terms of production and impact, the findings suggest an inverse relationship between publication volume and their impact per publication, and the divergence may not be entirely explained by ordinary citation-accumulation lag. On thematic composition, the findings suggest that the field's vocabulary has witnessed a certain shift towards GenAI-specific terms, with ChatGPT and LLM overtaking the previously dominant machine learning in terms of annual mention frequency starting from 2024. However, this shift was not strictly downwards in terms of promissory or speculative language: a sub-group of works used specific GenAI tools in combination with specific accuracy and readability assessments, pointing towards actual practical use. The field may be becoming more prolific in terms of GenAI-specific research, but it may be becoming less prolific in terms of research depth and breadth.

These two findings cannot dismiss a simple verdict of either evidence-generating research or publication growth, if only because the field is not characterised by an abundance of unsubstantial commentary. The presence of a genuine tool-evaluation cluster, as well as the absence of the type of runaway editorial proliferation reported in some of the other AI-health studies, may contribute to that conclusion. But neither is the field's rapid growth matched by a corresponding acceleration in the maturation of its evidentiary core, as suggested both by the decrease in the number of papers citing any given paper after 2023 and the peripheral role of the most GenAI-specific cluster. The most defensible answer to the question posed by this study is that the field's research output since ChatGPT's appearance is neither purely evidence-driven nor purely hype-driven but rather appears to be in a prolonged transitional state, one in which substantial conceptual and empirical work has begun but not yet completed its dialectical encounter with the eruptive growth of the field.

5.1. Limitations of the Study

Several limitations could be noted for this study. The present research applied database-assigned document types and bibliometric structural indicators, citation trajectory, thematic density and centrality, and keyword co-occurrence, as proxies for evidentiary composition, rather than manually or semi-automatically evaluating individual documents on their design type. This approach was adopted to permit analysis at full corpus scale (n = 4,358) across a merged, three-database dataset, a scale at which manual quality appraisal of every included record was not feasible within the scope of this study. Consequently, the study cannot distinguish, for example, a randomised controlled trial from a narrative review published under the same document-type label, and the findings regarding methodological rigor should be read as bibliometric-level patterns

suggestive of, rather than direct confirmation of, the evidentiary composition of the underlying literature. The citation-based findings are also limited by the absence of citations in PubMed, so citation-based evidence was extracted solely from WoS and Scopus parts of the combined database. Furthermore, the most recent in the field, notably the ones partially comprising the year 2026, were found to have limited impact evidence, being either too recent to have generated any impact yet or anticipated to experience delayed growth. Thus, the projection for the entire timeframe might differ from the actual findings. However, this study managed to provide the first-ever bibliometric analysis that sought to separate evidence from opinion in the claims regarding the rise of AI-driven health education in the post-ChatGPT age, providing a quantitative basis for discussion of the phenomenon. The tentative findings suggest that the field is still in the process of determining its future.

5.2. Implications for Policy and Practice

When evaluating research production in the field of AI and health education, the production and impact trajectory suggests that policymakers, funding agencies, and research councils should recognise that the volume of research is only a part of the picture. As this study demonstrates, the field experiences a disconnect between growth in research production and impact, with the latter being significantly lower than the former would suggest. For a policymaker or a funding agency considering an investment in the field, a significant year-over-year increase in published papers would be an indicator of relative success, but the observed decrease in impact would be informative of the field's needs. Funding and evaluation agencies responsible for this field should assign higher importance in the assessment of grant portfolios or national science priorities in this domain by considering field-normalised, time-adjusted citation metrics, or better, direct evidence of downstream educational or clinical outcome validation, more heavily than raw output counts.

In addition, the evidence that GenAI-specific terminology is both prevalent and peripheral to the overall structure of the literature suggests that quantitative analysis of publication volumes on ChatGPT or other LLMs is not an appropriate metric for determining the maturity of evidence in the field. A funder evaluating this literature landscape based on this metric alone would be misled to believe the field is significantly more developed than is the case. Indeed, the thematic-map result suggests that while evaluative capacity does exist, the evidence is not yet consolidated, a consideration for future health-related government agencies such as the WHO (2024), whose own guidance on large multi-modal models in healthcare was published at a similar inflection point for the field. Discussing the emergence of GenAI in healthcare from an implementation science perspective, Reddy (2024) highlighted the importance of developing frameworks for evidence-generation around application, integration, and governance, rather than relying on publication volumes as a proxy for readiness, a position that resonates directly with the findings of this thematic map.

International and national health governance agencies should consider the findings of this study alongside the WHO (2024) guidance on large multi-modal models in health. The guidance is a timely, high-profile statement from an international regulatory body, and the disconnect between the growth in research production and impact found by this study suggests that actionable plans should be directed toward balancing the rapid rise in literature production with its quality by policymakers, not merely a single guidance release.

When reviewing the literature for possible adoption of AI tools and methodologies in their practice, health educators and curriculum developers should be aware that the production and impact trajectory shows that a higher number of publications on a particular topic does not equal a higher level of validation or readiness for use. Specific attention should be paid to technological relevance, as the papers published before 2023, including many review, commentary, and narrative articles, may well have used older technology no longer relevant to the practice. Additionally, if considering the recommendations of a systematic review or a narrative literature review, users should be aware that the findings of this study suggest that reviews published in 2023, right after

ChatGPT's launch, may have included a significantly greater proportion of unsubstantiated commentary pieces than later reviews.

Another practical takeaway for health educators and curriculum developers is specific rather than generic; the evaluative sub-literature identified in the factorial analysis, work explicitly pairing named GenAI tools with accuracy and readability assessment, is the preferred evidence base for adoption decisions over the much larger set of GenAI-adjacent literature generally. Practitioners searching this literature should accordingly prioritise retrieval strategies seeking this evaluative cluster specifically (e.g., combinations of particular tools with accuracy, readability, and validation terms) over the field's dominant but thematically undifferentiated keywords, including "artificial intelligence" and "ChatGPT" generally.

When evaluating a large-scale recommendation to adopt a technology based on a literature review, institutional adoption committees considering this study's findings should be aware that unsubstantiated review pieces dominate the literature on the use of GenAI in health education. This being the case, such an institutional committee may want to advocate for a wait-and-see approach and require evidence from a controlled trial before implementing any large-scale methodological shifts proposed by the literature.

According to Naqvi et al. (2025) on critical thinking in health sciences education in the age of GenAI, health-education curricula should incorporate teaching critical-appraisal competencies related to GenAI-specific content into their training, rather than relying on such tools as instructional aids only. The finding of the current study on GenAI-specific thematic clusters can be seen as supporting that recommendation, as it highlights the relative lack of research on GenAI-specific issues in comparison to the general topic of health/science education. If the research field has not yet matured, then health-education professionals and students need to rely on their own critical appraisal skills to process information that comes from GenAI, rather than deferring to a body of literature that is still consolidating.

5.3. Future Research Agenda

Several avenues of future research would benefit from this study's findings without replicating it. A comparison between AI in the health education field-normalised citation trends and similar fields such as biomedical subfields in their emergent stages would be informative: this study's findings on the volume-impact disconnect would complement the findings of a similar study performed on another area of health scholarship or medicine. Such a study would be useful to determine whether the trends observed in this study are unique to the field of health education or apply to other areas of scientific literature as well. The field of schizophrenia genetics, which is currently experiencing a rapid upsurge in literature production post-2022, would be a viable candidate for such a study, as the spike appears to be caused by a similar set of factors unrelated to GenAI.

Future research analysing the trends observed in this study would benefit from stratifying the analysed documents by type. This study, while utilising databases' automated document-type detection, did not make use of this information beyond analysis of the overall curve. Specifically, future research would benefit from exploring whether the trends described in this study apply across different types of documents, or are driven primarily by commentaries and reviews, as suspected in this study. It would be informative to revisit this study's findings in three to five years, when the 2024, 2025, and particularly the 2026 outputs will have been out for a sufficient amount of time to experience citation-accumulation lag. It is suspected that some, but not all, of the trends described in this study would persist due to the evidence of this study's time-normalised citation count growth, but only further analysis will confirm this suspicion.

Since the WHO (2024) guidance specifically recommends post-release auditing of large multi-modal models in health by independent third parties, future research may productively explore the extent to which the published literature adheres to the recommendations, particularly the guidance's requirements related to reporting of the AI lifecycle. This study, while identifying the

need for such auditing, is unable to explore the topic in greater detail, but future research, particularly within a journal, may productively focus on analysing the extent to which the submissions follow the suggested reporting requirements, either as an overarching analysis or in greater detail in accordance with the requirements.

A sentiment and discourse-analytic study of the abstracts of AI-health-education studies following the methodology used by Tao and Shen (2025) for ChatGPT discourse in the social sciences would directly quantify the promissory versus evaluative tone at the level of abstracts and hence test the present findings of broad but shallow themes. This would be a more direct test of the hypothesis that the detected thematic pattern reflects genuine evaluative content beyond mere keyword clustering. A release-anchored temporal analysis identifying keyword trends associated with specific dates of GenAI model releases (GPT-4, GPT-4o, Gemini, DeepSeek), rather than ChatGPT's initial appearance in November 2022, can challenge Yoga Ratnam (2025) proposition on the proximity of subsequent capability releases to this literature's lexical trends. The findings of the current study appear to support the scholar's hypothesis, though not structured to directly test it.

A cross-domain comparative thematic mapping study, applying the same density-centrality method to the larger GenAI-in-healthcare literature that Moulaei et al. (2024) scoped narratively, would determine whether the thematic immaturity identified in this study was specific to the health education domain or characteristic of a larger body of work, a distinction with downstream implications for whether the recommendations above should be specified to health-education governance or apply more broadly to the research program on GenAI in healthcare. A prospective re-run of the present study's thematic mapping in three to five years, once the large language model/chatgpt/patient education cluster has had time to consolidate its position relative to the field's core themes, would identify whether the present study's most striking finding, the predominance of volume over consolidation was primarily a function of the cluster's developmental timing or a feature of its structural relationship to the rest of the field.

To move this study forward, it is explicitly recommended that future work incorporate manual or AI-assisted design-type classification of a representative subsample, citing this as the natural and necessary next step to move from the exploratory, pattern-level findings this study offers toward a direct, confirmatory test of the evidence-generating research or publication growth question.

References

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., Al Yami, M. S., Al Harbi, S., & Albekairy, A. M. (2023). Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Medical Education, 23*(1), Article 689. https://doi.org/10.1186/s12909-023-04698-z

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

Aria, M., Cuccurullo, C., D'Aniello, L., & Spano, M. (2026). Biblioshiny and the SAAS workflow: An integrated framework for transparent and reproducible science mapping. A demonstration through the replication of a study. *Journal of Informetrics, 20*(3), Article 101837. https://doi.org/10.1016/j.joi.2026.101837

Arzilli, G., Di Maggio, E., De Angelis, L., Baglivo, F., Savoia, E., Privitera, G. P., & Rizzo, C. (2025). A surge of AI-driven publications: The impact on health professionals and potential mitigating solutions. *Frontiers in Public Health, 13*, Article 1680630. https://doi.org/10.3389/fpubh.2025.1680630

Caputo, A., & Kargina, M. (2022). A user-friendly method to merge Scopus and Web of Science data during bibliometric analysis. *Journal of Marketing Analytics, 10*(1), 82–88. https://doi.org/10.1057/s41270-021-00142-7

Cheng, K., Li, Z., Sun, Z., Guo, Q., Li, W., Lu, Y., Qi, S., Shen, Z., Xie, R., Wang, Y., Wu, Z., Wu, Y., Wu, C., Li, Y., Xie, Y., Wu, H., & Li, C. (2024). The rapid growth of bibliometric studies: A call for international guidelines. *International Journal of Surgery, 110*(4), 2446–2448. https://doi.org/10.1097/JS9.0000000000001049

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

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal, 22*(2), 338–342. https://doi.org/10.1096/fj.07-9492LSF

Genc, H. N., & Kocak, N. (2025). Bibliometric analysis of studies on artificial intelligence in environmental and health education. *Journal of Education in Science, Environment and Health, 11*(2), 140–150. https://doi.org/10.55549/jeseh.806

Guo, Y., Guo, Q., Liu, C., Chen, F., Ma, J., Zhang, H., Lv, Y., Yang, T., & Sun, Y. (2026). Artificial intelligence in patient education: A bibliometric analysis. *International Journal of Surgery, 112*(3), 6134–6149. https://doi.org/10.1097/JS9.0000000000004075

Iqbal, U., Tanweer, A., Rahmanti, A. R., Greenfield, D., Lee, L. T.-J., & Li, Y.-C. J. (2025). Impact of large language model (ChatGPT) in healthcare: An umbrella review and evidence synthesis. *Journal of Biomedical Science, 32*(1), Article 45. https://doi.org/10.1186/s12929-025-01131-z

Moulaei, K., Yadegari, A., Baharestani, M., Farzanbakhsh, S., Sabet, B., & Afrash, M. R. (2024). Generative artificial intelligence in healthcare: A scoping review on benefits, challenges and applications. *International Journal of Medical Informatics, 188*, Article 105474. https://doi.org/10.1016/j.ijmedinf.2024.105474

Nikolić, D., Ivanović, D., & Ivanović, L. (2024). An open-source tool for merging data from multiple citation databases. *Scientometrics, 129*(7), 4573–4595. https://doi.org/10.1007/s11192-024-05076-2

Naqvi WM, Ganjoo R, Rowe M, Pashine AA and Mishra GV (2025) Critical thinking in the age of generative AI: implications for health sciences education. Front. Artif. Intell. 8:1571527. doi: 10.3389/frai.2025.1571527

Ningrum, D. N. A., & Muhtar, M. S. (2024). A bibliometric analysis of artificial intelligence research in public health education. In N. A. S. Abdullah, T. S. Hoon, N. M. Shamsudin, & R. Legino (Eds.), *Proceedings of the International Conference on Innovation & Entrepreneurship in Computing, Engineering & Science Education (InvENT 2024)* (Vol. 117, pp. 298–307). Atlantis Press International B.V. https://doi.org/10.2991/978-94-6463-589-8_27

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Paleti, S. T., Kambhampati, S. B. S., Vaish, A., Vaishya, R., & D'Ambrosi, R. (2025). Rise of Asian research in orthopaedic and sports medicine: A bibliometric analysis from 1996 to 2022. *European Journal of Orthopaedic Surgery & Traumatology, 35*(1), Article 173. https://doi.org/10.1007/s00590-025-04294-5

Reddy, S. (2024). Generative AI in healthcare: An implementation science informed translational path on application, integration and governance. *Implementation Science, 19*(1), Article 27. https://doi.org/10.1186/s13012-024-01357-9

Sallam, M., Barakat, M., & Sallam, M. (2024). A preliminary checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence–based models in

health care education and practice: Development study involving a literature review. *Interactive Journal of Medical Research, 13*, Article e54704. https://doi.org/10.2196/54704

Tao, Y., & Shen, Q. (2025). Academic discourse on ChatGPT in social sciences: A topic modeling and sentiment analysis of research article abstracts. *PLOS ONE, 20*(10), Article e0334331. https://doi.org/10.1371/journal.pone.0334331

Temsah, M.-H., Altamimi, I., Jamal, A., Alhasan, K., & Al-Eyadhy, A. (2023). ChatGPT surpasses 1000 publications on PubMed: Envisioning the road ahead. *Cureus, 15*(9), Article e44769. https://doi.org/10.7759/cureus.44769

World Health Organization. (2024). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. https://iris.who.int/handle/10665/375579

Yoga Ratnam, K. K. (2025). Generative artificial intelligence in public health research and scientific communication: A narrative review of real applications and future directions. *Digital Health, 11*, Article 20552076251362070. https://doi.org/10.1177/20552076251362070