

Research Trends and Knowledge Structure of Large Language Models in Medical Ethics: A Bibliometric Analysis

Bo Wang¹, Syed Ahmed Salman^{1*}, and Oyyappan Duraipandi^{1*}

¹School of Business & Management, Lincoln University College, 47301 Selangor, Malaysia

*Corresponding author: Oyyappan@lincoln.edu.my

*Co-corresponding author: syedahmed@lincoln.edu.my

Abstract

Large language models (LLMs) are increasingly being applied in healthcare and medical education, raising growing ethical concerns regarding reliability, accountability, bias, and governance. However, the global research landscape and thematic structure of studies addressing LLM-related medical ethics remain insufficiently characterized. A bibliometric analysis was conducted using publications indexed in the WoSCC. Relevant studies published up to December 31, 2025 were retrieved and screened according to predefined eligibility criteria. Bibliometric mapping was performed using VOSviewer and the Bibliometrix R package to analyze publication trends, collaboration networks, author keyword co-occurrence, and interdisciplinary citation structures. A total of 1,604 publications were included in the final analysis. The annual number of publications increased markedly after 2022, coinciding with the rapid expansion of generative AI technologies. Sixty-five countries or regions contributed to this research field, with the United States producing the largest number of publications and occupying the most central position in the collaboration network. Keyword co-occurrence analysis identified five major thematic clusters, including clinical and educational applications of LLMs, ethical governance and regulation, methodological foundations of AI in healthcare, digital mental health applications, and core bioethical principles. Citation flow analysis revealed strong interdisciplinary knowledge exchange between healthcare sciences and computer science, highlighting the hybrid nature of this emerging research domain. Research on LLMs and medical ethics is expanding rapidly and demonstrates increasing interdisciplinary integration. Future studies should focus on developing robust ethical frameworks, governance strategies, and empirical evaluations to support the responsible implementation of LLM technologies in healthcare.

Index Terms— large language models, medical ethics, generative artificial intelligence, healthcare, bibliometric

1 Introduction

Large language models (LLMs), built upon transformer-based architectures and trained on massive corpora, have rapidly

transformed the landscape of artificial intelligence (Achille et al., 2024). Following the public release of conversational systems such as ChatGPT in late 2022, generative AI technologies have entered widespread clinical, academic, and public discourse (Yu et al., 2025). Unlike earlier task-specific machine learning systems, LLMs demonstrate broad linguistic competence, contextual reasoning, and adaptability across domains, thereby accelerating their integration into knowledge-intensive fields such as medicine (Wu et al., 2025).

In healthcare, LLMs have been applied to diverse tasks including clinical decision support, medical documentation, patient education, diagnostic assistance, and research synthesis (Sandmann et al., 2025; Seo et al., 2024). Their deployment extends across specialties ranging from radiology and oncology to psychiatry and primary care (Bhayana, 2024; Caglayan et al., 2024; Volkmer et al., 2024). Simultaneously, medical education institutions have incorporated LLM-based tools into training environments, reshaping pedagogical practices and professional skill development (Lucas et al., 2024). This rapid expansion has positioned LLMs not merely as technical instruments but as socio-technical systems embedded within healthcare infrastructures.

Alongside their growing adoption, LLMs have raised substantial ethical and governance-related concerns. Issues such as algorithmic bias, misinformation, data privacy, accountability, transparency, patient autonomy, and the reliability of AI-assisted decision-making have become central topics of scholarly debate (Barr et al., 2025; Lin et al., 2024; Ong et al., 2024). In high-stakes clinical contexts, inaccuracies or opaque model outputs may have direct implications for patient safety and professional responsibility (Trujillo et al., n.d.). Moreover, the use of conversational AI in sensitive areas such as mental health care introduces additional considerations related to vulnerability, trust, and human-machine interaction (L. Wang et al., 2025). As a result, discussions surrounding LLMs increasingly intersect with established principles of medical ethics and emerging regulatory frameworks.

Although the ethical implications of artificial intelligence in healthcare have been examined for more than a decade, research specifically addressing large language models repre-

Running title: LLMs and Medical Ethics: A Bibliometric Study

Oyyappan Duraipandi, ORCID: 0000-0001-6076-0720.

Syed Ahmed Salman, ORCID: 0000-0001-7343-2064.

sents a rapidly evolving and highly dynamic subfield. The accelerated growth of publications since 2023 suggests that LLM-related scholarship is expanding at a pace that may outstrip conceptual consolidation. To date, however, a comprehensive mapping of the intellectual structure, thematic clusters, collaborative patterns, and evolving ethical priorities within this field remains limited. To address this gap, the present study conducts a large-scale bibliometric analysis of publications indexed in the Web of Science Core Collection. By examining publication trends, collaboration networks, keyword co-occurrence structures, and thematic evolution, this study aims to systematically characterize the knowledge architecture of research on large language models in medical contexts and to identify the principal ethical dimensions shaping this emerging field.

2 Methods

2.1 Data Source and Search Strategy

All bibliographic data were obtained from the Web of Science Core Collection (WoSCC), which was selected for its rigorous indexing standards and suitability for citation analysis.

A comprehensive search was conducted on February 1, 2026 by using the following strategy:

```
TS=( ("large language model*" OR LLM* OR ChatGPT OR
GPT-4 OR GPT4 OR GPT-3 OR GPT3 OR "generative AI" OR
"foundation model*" OR "transformer model*" OR LLaMA OR
PaLM OR Gemini OR Claude)
AND (medic* OR health* OR clinical)
AND (ethic* OR moral* OR governance OR regulation OR
accountability OR liability OR "informed consent") )
AND PY=(2001-2025)
```

2.2 Eligibility Criteria

Inclusion criteria:

1. Publications focusing on large language models, generative artificial intelligence, or related technologies in the context of medical ethics, healthcare ethics, or ethical implications in medicine.
2. English articles indexed in the Web of Science Core Collection.
3. Original research articles or review articles.

Exclusion criteria:

1. Publications not directly related to the ethical, clinical, or governance implications of large language models in healthcare.
2. Editorials, letters, meeting abstracts, corrections, news items, or other non-research documents.
3. Duplicate records identified during the data cleaning process.

2.3 Data Extraction and Processing

The full records were exported in text format. Duplicates were identified and removed using R (version 4.5.1) and manual cross-checking. To address inconsistencies (e.g., missing keywords or abbreviations), we performed data cleaning and harmonization (e.g., merging “AI” and “artificial intelligence”).

2.4 Bibliometric and Visualization Analysis

The bibliometric analysis was conducted by using VOSviewer (v1.6.20) and Bibliometrix R-package (v4.5.1). The following analyses were performed: Publication trends and general characteristics were examined through annual output, country or region contributions, and Web of Science subject categories. Collaboration structures were analyzed using co-authorship networks at the author, institutional, and country levels. Research themes were identified through author keyword co-occurrence clustering. Interdisciplinary knowledge flows were explored using dual-map overlay analysis, citation flow matrices, and disciplinary diversity indices.

2.5 Study Selection

A total of 2,634 records were initially retrieved from the Web of Science Core Collection. After restricting the dataset to English-language publications, articles or review articles, and publications indexed up to December 31, 2025, 2,180 records remained. Subsequently, two independent reviewers screened the records according to the predefined eligibility criteria by examining titles, keywords, and abstracts. When necessary, full texts were reviewed to determine relevance. During this stage, 576 records were excluded because they were not directly related to the ethical implications of large language models or generative AI in healthcare, leaving 1,604 studies for the final bibliometric analysis (Figure 1).

3 Results

3.1 Publication Trends and General Characteristics

The annual number of publications related to large language models (LLMs) and medical ethics has increased substantially in recent years (Figure 2). Early publications were limited in number, whereas a pronounced growth in output occurred following the emergence and widespread adoption of generative AI technologies. The total publication curve shows a clear upward trend across the study period.

The country or region distribution demonstrates that research output is geographically diverse, with 65 countries or regions participating in this field. Based on fractional publication counts, the United States produced the largest number of publications (590), followed by China (217), England (152), Germany (111), Australia (103), India (96), Italy (95), and Canada (84).

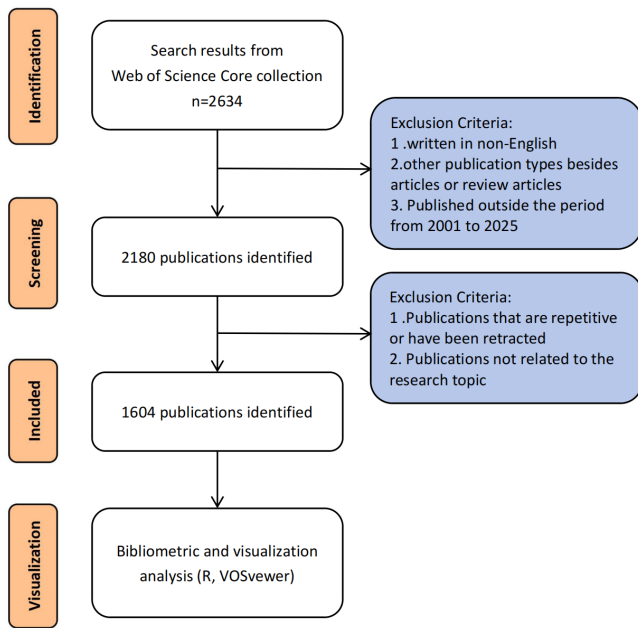


Figure 1. Flowchart of the search strategy used in this study

In addition to publication volume, the collaboration network reveals varying levels of international connectivity. The United States occupies the most central position in the network, exhibiting the highest total link strength (476) and maintaining collaboration links with a large number of countries. Other countries such as England, Australia, China, Italy, and Canada also display relatively high levels of collaborative connectivity.

The disciplinary distribution of publications further demonstrates the multidisciplinary nature of the field (Figure 3). The most represented Web of Science subject categories include Health Care Sciences & Services, Computer Science, General & Internal Medicine, Medical Informatics, and Education & Educational Research. Additional categories such as Psychology, Oncology, and Public, Environmental & Occupational Health also appear among the top subject areas.

Overall, the distribution of publications across countries and subject categories indicates that research on LLMs and medical ethics spans multiple disciplinary domains while involving broad international participation.

3.2 Global and Institutional Collaboration Structure

The co-authorship network analysis reveals collaboration patterns at the author, institutional, and country levels (Figure 4A–C).

At the author level, the network includes 169 authors who published at least three articles in the field. Several authors occupy relatively central positions within the collaboration network. Among them, Wisit Cheungpasitporn (10 publications), Charat Thongprayoon (9), and Zohar Elyoseph (8) represent the most productive contributors, while John Torous and Dominic Wilkinson each authored 7 publications. These

authors form several collaborative clusters connected through co-authored publications, indicating the presence of multiple research teams working on related topics.

At the institutional level, the collaboration network consists of 126 organizations with at least five publications. A number of institutions with higher publication outputs function as key nodes in the network and maintain repeated collaborative relationships with other organizations. Institutional collaboration clusters are formed through co-authored publications, reflecting structured research partnerships across universities, hospitals, and research centers.

At the country or region level, the collaboration network includes 65 countries or regions participating in the field. The United States occupies the most central position in the network, producing 590 publications and exhibiting the highest total link strength (476). Other countries with substantial publication outputs include China (217 publications), England (152), Germany (111), Australia (103), India (96), Italy (95), and Canada (84). Several of these countries also display strong collaborative connections with multiple partners, forming the major hubs of the global research network.

The annual distribution of collaboration patterns further illustrates the dynamics of international cooperation (Figure 4D). The total number of publications increased markedly from 2022 to 2025, accompanied by a steady rise in both single-country and multi-country publications. Across all years, single-country publications constituted the majority of research output, while international collaborations accounted for a smaller but substantial share of publications. The international collaboration rate fluctuated within a relatively stable range, increasing from 28.7% in 2022 to 31.3% in 2023 and reaching approximately 35% in 2024, before remaining at a comparable level in 2025. These patterns indicate a gradual expansion of international research cooperation alongside the overall growth of the field.

3.3 Thematic Clustering of Author Keywords

The author keyword co-occurrence network consisted of 141 nodes, 1,712 links and total link strength 6,813, forming five clusters (Figure 5; Table 1). The network displayed a centralized structure, with Cluster 1 occupying the largest share of nodes and total link strength. The strongest inter-cluster connections were observed between Cluster 1 and Cluster 2, followed by Cluster 1 and Cluster 3.

Cluster 1: Clinical and Educational Applications of LLMs

Cluster 1 represents the largest group (50 keywords; total occurrences = 2,597). It is centered on large language model, chatgpt, and artificial intelligence, with high-frequency terms including medical education, healthcare, clinical decision support, patient education, and medicine. The keywords in this cluster form dense internal linkages and maintain the highest number of cross-cluster connections, particularly with Clusters 2 and 3. The clustering pattern indicates that research on LLMs is strongly positioned within medical training en-

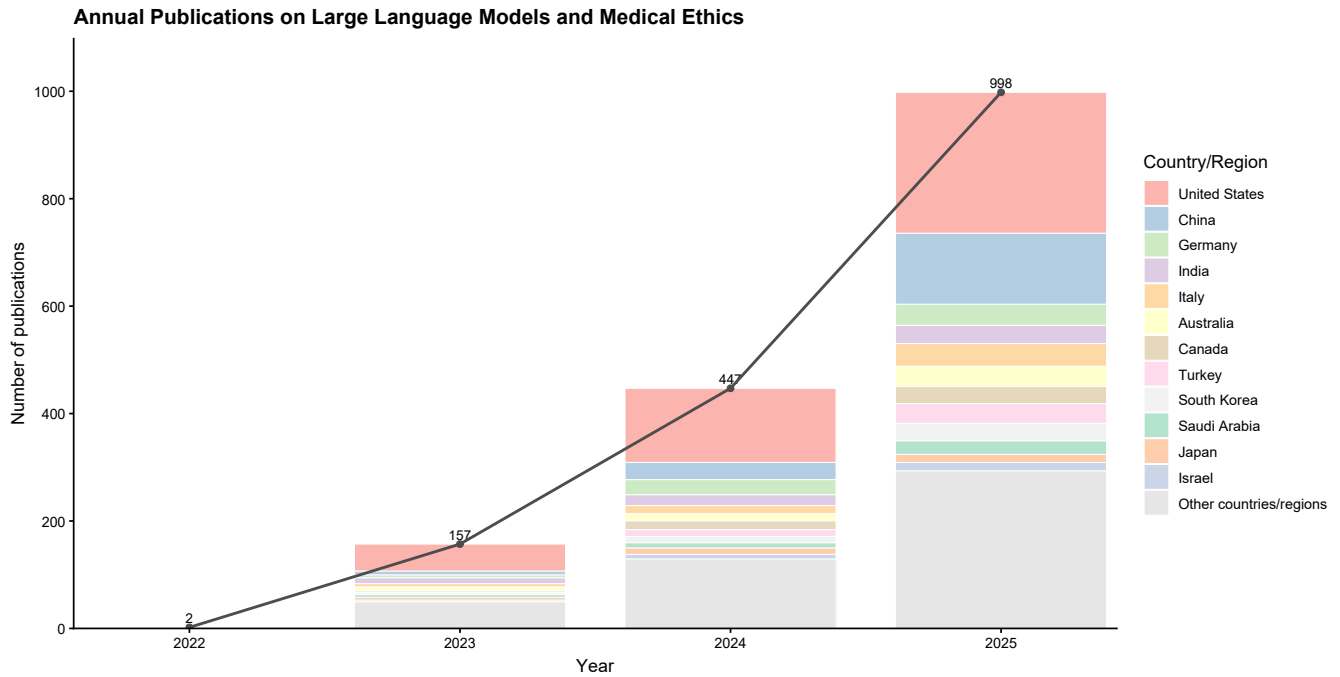


Figure 2. Annual publications by country or region

The stacked bar chart presents the annual number of publications on large language models and medical ethics by country or region. Each publication was counted once for every country or region listed in the author affiliations. Bars represent country-level publication counts, while the overlaid line indicates the total number of publications per year.

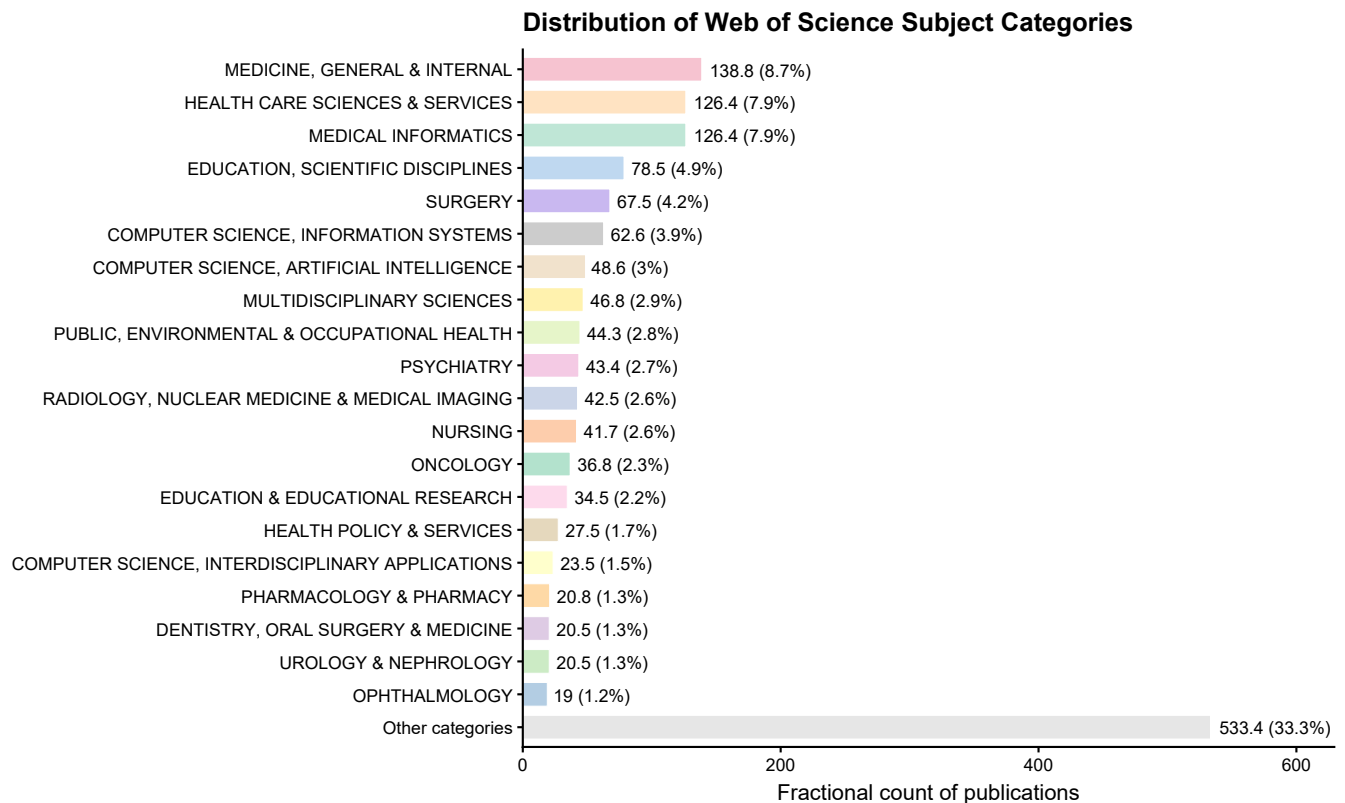


Figure 3. Distribution of Web of Science subject categories.

The horizontal bar chart presents the top 20 Web of Science subject categories represented in the dataset. Fractional counting was applied to account for multiple category assignments per publication. Bars indicate the fractional number of publications in each category, and percentages represent the proportion relative to the total number of publications.

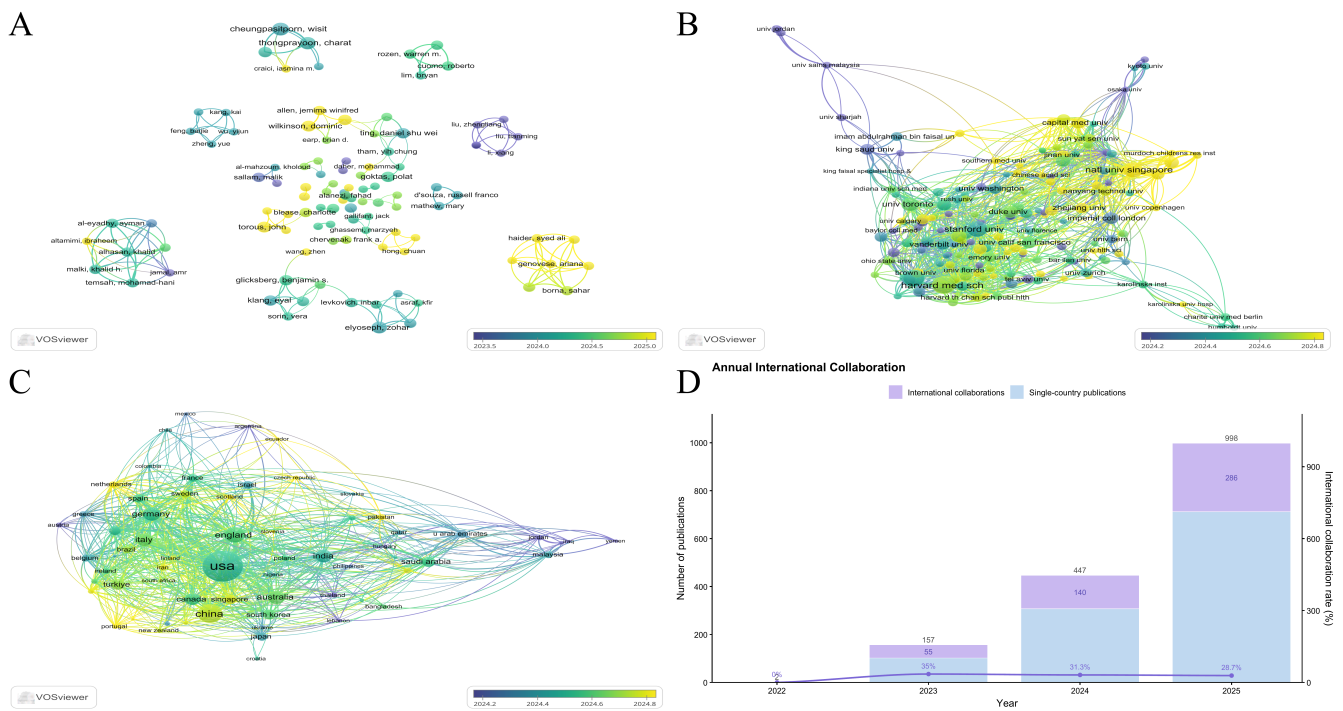


Figure 4. Multi-level collaboration structure and international collaboration dynamics

Panels (A–C) present co-authorship networks at three analytical levels using VOSviewer.

(A) Author-level network including authors with ≥ 3 publications.

(B) Organizational-level network including institutions with ≥ 5 publications.

(C) Country/region-level network including countries/regions with ≥ 5 publications.

In all network maps, node size is proportional to publication output, and link thickness reflects the strength of collaborative relationships based on co-authored publications. Nodes are grouped into clusters according to VOSviewer's modularity-based clustering algorithm, with different colors indicating distinct collaboration communities.

Panel (D) illustrates annual international collaboration patterns. Stacked bars represent the yearly distribution of single-country and multi-country publications, while the overlaid line denotes the international collaboration rate (%). International collaboration was defined as publications involving affiliations from two or more countries/regions.

vironments and clinical decision-making workflows, where implementation-related discussions frequently intersect with questions concerning professional responsibility and safe integration into healthcare systems.

Cluster 2: Ethical Governance and Regulatory Issues of Generative AI

Cluster 2 contains 40 keywords (total occurrences = 806) and is organized around generative ai and ethics. Closely connected terms include artificial intelligence ethics, ethical concerns, data privacy, bias, regulation, privacy, and accuracy. This cluster demonstrates strong internal connectivity and substantial cross-links with Cluster 1. The keyword relationships are concentrated around governance-related and regulatory terminology.

Cluster 3: AI Methodological Foundations in Healthcare

Cluster 3 includes 24 keywords (total occurrences = 522) and comprises technical and methodological terms such as machine learning, natural language processing, deep learning, radiology, diagnosis, and precision medicine. This cluster sit-

uates LLM-related research within the broader methodological landscape of medical artificial intelligence, where issues of model performance, diagnostic accuracy, and system validation form part of the broader safety and accountability context in healthcare applications. Cluster 3 maintains strong linkages with Cluster 1 and moderate connections with Cluster 2.

Cluster 4: Digital Mental Health and Conversational Agents

Cluster 4 consists of 16 keywords (total occurrences = 344), characterized by chatbot, mental health, psychiatry, depression, digital health, and psychotherapy. The keyword configuration highlights the growing use of conversational AI in psychological and public health settings, areas in which the integration of LLM-driven systems occurs alongside discussions of patient interaction, vulnerability, and responsible digital care. This cluster maintains multiple links with Cluster 1 and fewer connections with the remaining clusters.

Cluster 5: Core Bioethical Principles in AI-Mediated Care

Cluster 5 comprises 11 keywords (total occurrences = 128), including medical ethics, informed consent, patient autonomy, confidentiality, and explainable ai. Compared with other clusters, Cluster 5 contains fewer nodes and lower total link strength. Its keywords are primarily related to foundational ethical concepts and exhibit limited cross-cluster connections. This cluster reflects the incorporation of foundational ethical principles into research examining the clinical implementation of large language models in healthcare.

3.4 Interdisciplinary Knowledge Flows and Citation Structure

The dual-map overlay of cross-disciplinary citation flows illustrates the patterns of knowledge exchange between disciplines (Figure 6). The left side of the map represents the disciplines of the citing publications, while the right side corresponds to the subject categories of the cited journals. Several prominent citation pathways are observed between medical, computational, and interdisciplinary science domains, indicating active knowledge exchange across these research areas.

The disciplinary publication flow analysis further shows how research outputs are distributed across journals (Figure 7). Publications originating from disciplines such as Health Care Sciences & Services, Computer Science, and General & Internal Medicine are linked to a group of frequently represented journals, suggesting that the field is characterized by both clinical and technological publication outlets.

The cross-disciplinary citation flow matrix provides a quantitative representation of these relationships (Table 2). Among the top disciplines, the largest citation flow was observed from Health Care Sciences & Services to Computer Science (180 links), followed by citations from Health Care Sciences & Services to General & Internal Medicine (90 links) and from Computer Science to Science & Technology – Other Topics (233 links). Additional notable citation relationships include flows from Education & Educational Research to Computer Science (52 links) and from Psychology to Computer Science (21 links), highlighting the methodological influence of computational approaches across multiple medical and social science domains.

The disciplinary diversity analysis further characterizes the breadth of interdisciplinary citation patterns (Table 3). Disciplines such as Dentistry, Oral Surgery & Medicine ($H = 1.3286$; normalized $H = 0.9584$), Psychology ($H = 1.5349$; normalized $H = 0.9537$), and Medical Informatics ($H = 1.5229$; normalized $H = 0.9463$) exhibit the highest normalized Shannon diversity values, indicating relatively balanced citation distributions across multiple target disciplines. In contrast, larger fields such as Computer Science (normalized $H = 0.6866$) and General & Internal Medicine (normalized $H = 0.6867$) show lower normalized diversity scores, suggesting more concentrated citation patterns despite their substantial overall citation volumes.

4 Discussion

The present bibliometric analysis reveals the rapid emergence of research at the intersection of large language models (LLMs) and medical ethics, while also illustrating how this field is being shaped by technological developments, disciplinary structures, and global research collaboration.

A first notable finding is the extraordinary acceleration of publication output, particularly after 2022. This temporal pattern strongly corresponds to the widespread release and adoption of generative AI systems, which rapidly expanded the practical possibilities of applying LLMs in healthcare (Moreno et al., 2023). The surge in publications therefore reflects not only growing academic interest but also the increasing urgency of addressing ethical implications associated with the clinical and educational deployment of generative AI (Al-kfairy et al., 2024a; Dave et al., 2023). In contrast to earlier phases of medical artificial intelligence research—where ethical discussions often emerged after technological applications were established—the development of LLMs appears to have triggered parallel growth in both technological exploration and ethical reflection (Moreno et al., 2023; Sallam, 2023). This simultaneity suggests that ethical governance has become an integral component of AI discourse in healthcare rather than a secondary consideration (Sung, 2023).

The geographic distribution and collaboration patterns further illustrate how this emerging research area is structured globally. Although 65 countries or regions contributed publications, the research network is strongly anchored by a group of technologically and scientifically advanced countries. The United States occupies a particularly central position, both in terms of publication output and international collaboration links (Ferrara, 2023; Gao et al., 2023). This dominance likely reflects its leading role in AI development, biomedical research infrastructure, and digital health innovation. At the same time, the presence of multiple additional hubs—including China, England, Germany, Australia, and Canada—indicates that research on LLM-related ethics is developing within a multi-regional knowledge network rather than a single dominant research center (Casella et al., 2023; Dave et al., 2023; Le-Khac & Truong, 2025; Moreno et al., 2023). The gradual increase in international collaboration rates also suggests that ethical questions surrounding AI deployment are increasingly being addressed through cross-national scientific dialogue, reflecting the global nature of both digital technologies and healthcare governance.

The thematic structure of the field provides further insight into how research priorities are evolving. The keyword clustering results indicate that the intellectual landscape of LLM-related medical ethics research is organized around three interconnected layers: practical implementation, ethical governance, and technological foundations (Jia & Zhao, 2025; Workum et al., 2025; Zhui et al., 2024). The largest cluster focuses on clinical and educational applications of LLMs, highlighting how generative AI is being incorporated into medical education, clinical decision support, and healthcare communication (Boscardin et al., 2024; Zohny et al., 2025). The cen-

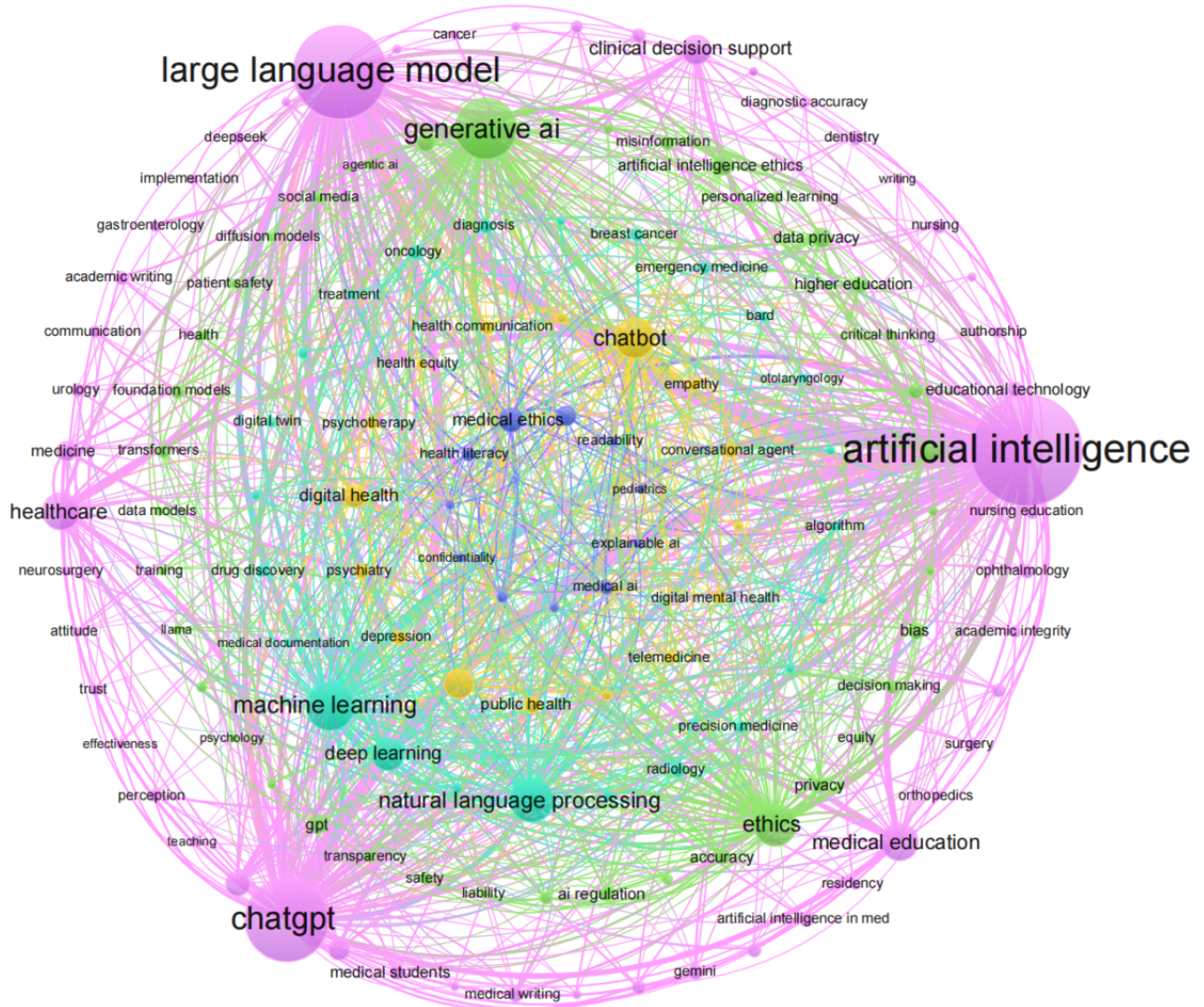


Figure 5. Keyword co-occurrence network and cluster structure generated by VOSviewer.

The keyword co-occurrence network was constructed based on author keywords extracted from the WoScc and visualized using VOSviewer. Each node represents a keyword, and node size is proportional to its occurrence frequency. Links between nodes indicate co-occurrence relationships within the same publication, with link thickness reflecting the total link strength. Nodes with the same color belong to the same cluster, representing internally cohesive research themes. The spatial proximity between nodes reflects the strength of their association, with closely positioned keywords indicating stronger co-occurrence relationships. The resulting cluster structure reveals the major thematic domains and the knowledge organization of the research field.

tral position of this cluster within the keyword network suggests that the field is currently driven largely by questions related to real-world implementation and practical use scenarios. Closely linked to this application-oriented research is a second major thematic domain addressing ethical governance and regulatory issues. Topics such as bias, data privacy, transparency, and AI regulation appear prominently, reflecting the growing recognition that generative AI systems introduce novel ethical challenges in healthcare contexts (Al-kairy et al., 2024b; Ferrara, 2023; Raman, 2025). The strong connections between the application cluster and the governance cluster indicate that ethical concerns are not isolated theoretical debates but instead arise directly from the practical deployment of LLM

systems. In other words, ethical governance in this field appears to develop in response to emerging implementation challenges, rather than existing as an independent conceptual research area (Maity & Saikia, 2025). A third thematic layer is represented by the cluster focusing on methodological foundations of AI in healthcare. This cluster includes core technological approaches such as machine learning, natural language processing, and predictive modeling (Alqahtani et al., 2023; Androulakis et al., 2025). Although these topics appear more technical, their connections with other clusters demonstrate that technological development remains closely intertwined with clinical and ethical considerations. This relationship reflects the fact that issues such as model accuracy, reliability,

Table 1. Co-occurrence clustering of author keywords

Cluster	Topic	Keywords
1	Clinical and Educational Applications of LLMs	artificial intelligence, large language model, chatgpt, healthcare, medical education, clinical decision support, education, patient education, medicine, educational technology, medical students, gemini, artificial intelligence in medicine, medical writing, medical research, academic writing, nursing education, ophthalmology, research ethics, clinical practice, patient care, nursing, dentistry, urology, surgery, perception, deepseek, academic integrity, dental education, communication, trust, authorship, gastroenterology, residency, attitude, orthopedics, implementation, diagnostic accuracy, nephrology, cancer, health informatics, neurosurgery, effectiveness, plastic surgery, writing, healthcare technology, nursing students, teaching, telehealth, research integrity
2	Ethical Governance and Regulatory Issues of Generative AI	generative ai, ethics, artificial intelligence ethics, ethical concerns, data privacy, prompt engineering, higher education, gpt, bias, privacy, accuracy, ai regulation, openai, decision making, foundation models, misinformation, transformers, safety, transparency, medical services, social media, reliability, health, retrieval-augmented generation, responsible ai, equity, synthetic data, personalized learning, medical diagnostic imaging, training, liability, patient safety, data models, healthcare professionals, critical thinking, diffusion models, digital literacy, llama, agentic ai, psychology
3	AI Methodological Foundations in Healthcare	machine learning, natural language processing, deep learning, radiology, breast cancer, oncology, diagnosis, emergency medicine, drug discovery, precision medicine, electronic health records, sentiment analysis, bard, reinforcement learning, digital twin, algorithm, treatment, predictive models, predictive analytics, otolaryngology, pathology, diagnostic imaging, medical documentation, personalized medicine
4	Digital Mental Health and Conversational Agents	chatbot, mental health, digital health, public health, conversational agent, psychiatry, health communication, depression, empathy, psychotherapy, digital mental health, health education, telemedicine, anxiety, health equity, human-computer interaction
5	Core Bioethical Principles in AI-Mediated Care	medical ethics, informed consent, medical ai, health literacy, information technology, explainable ai, readability, patient engagement, confidentiality, patient autonomy, pediatrics

Table 2. Core cross-disciplinary citation flow matrix among the top 10 disciplines.

Citing \ Cited	HCSS	CS	GIM	Oncol	STOT
Others	26	180	138	27	475
Health Care Sci & Services	0	117	90	14	220
Computer Science	17	0	79	15	233
General & Internal Medicine	10	37	0	6	138
Education & Educational Research	7	52	32	0	103
Oncology	0	14	16	0	78
Science & Technology – Other Topics	4	65	16	0	0
Psychology	8	21	13	0	21
Medical Informatics	13	15	13	0	33
Surgery	0	7	20	0	39

Note: Rows represent citing disciplines and columns represent cited disciplines. HCSS = Health Care Sciences & Services; CS = Computer Science; GIM = General & Internal Medicine; Oncol = Oncology; STOT = Science & Technology – Other Topics. Diagonal elements (within-discipline self-citations) were excluded. Only the top 10 disciplines ranked by total outgoing cross-disciplinary citation volume were retained. Cell values indicate the number of cross-disciplinary citation links between each disciplinary pair.

and algorithmic transparency are not only technical concerns but also central elements of ethical evaluation in AI-assisted medical decision-making (Meskó & Topol, 2023; Temizsoy Korkmaz et al., 2025; Xu et al., 2024). Two smaller clusters highlight emerging or specialized directions within the field. One focuses on conversational AI applications in digital mental health, reflecting the rapid expansion of chatbot-based sys-

tems in psychological support and public health communication (Guo et al., 2024). These applications raise unique ethical challenges related to patient vulnerability, emotional interaction, and responsibility in automated care environments (Tahir et al., 2025). The other cluster centers on traditional bioethical principles such as informed consent, patient autonomy, and confidentiality. Although this cluster is smaller, its pres-

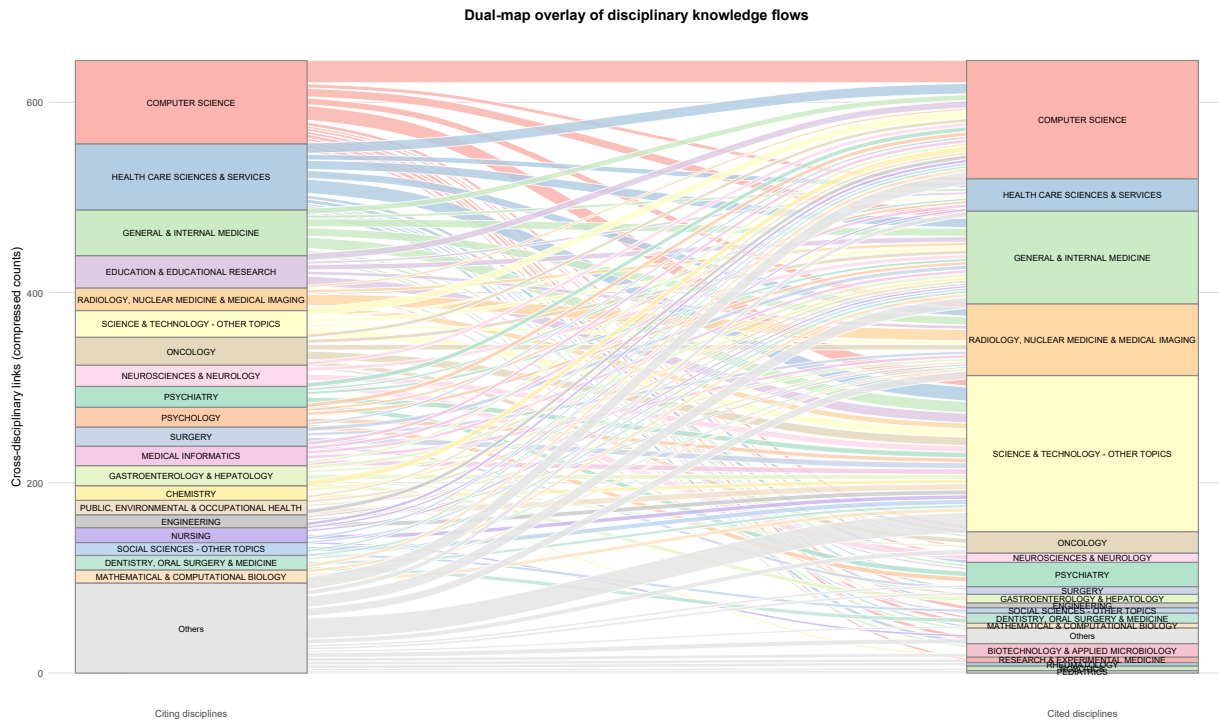


Figure 6. Dual-map overlay of disciplinary knowledge flows (Discipline → Discipline)

The left column represents the primary Web of Science subject categories (disciplines) of the citing publications, and the right column represents the subject categories of the cited journals derived from reference data. The width of each flow is proportional to the number of cross-disciplinary citation links (square-root scaled for visualization). To enhance interpretability, only the top 20 disciplines on each side are shown, with remaining categories aggregated into “Others.” The figure illustrates patterns of knowledge diffusion and interdisciplinary integration within the field, highlighting dominant citation pathways and the extent of cross-domain intellectual exchange.

ence indicates that foundational ethical frameworks continue to provide important conceptual guidance for evaluating new AI technologies in healthcare (Allen et al., 2025, 2026).

The interdisciplinary structure of the field further reinforces the interpretation that LLM-related medical ethics research is evolving as a hybrid domain bridging medicine, computing, and social sciences. The dual-map overlay and citation flow analysis reveal strong citation pathways between healthcare sciences and computer science, demonstrating that clinical research frequently draws upon computational methodologies (Liu, 2023; H. Wang et al., 2023). At the same time, flows toward broader science and technology domains indicate that the field is embedded within a wider technological research ecosystem. The disciplinary diversity analysis provides additional evidence of this interdisciplinary integration. Fields such as psychology and medical informatics exhibit relatively broad citation distributions across multiple disciplines, suggesting that they function as important integrative interfaces linking technological development with clinical and social dimensions of healthcare (Choudhury & Shamszare, 2023; De Freitas et al., 2024).

Taken together, these findings suggest that research on LLMs and medical ethics is currently transitioning from an early exploratory stage toward a more structured interdisciplinary field. The strong emphasis on practical applications indicates that technological adoption is occurring rapidly, while the simultaneous emergence of governance-related research

reflects growing awareness of the ethical implications of AI-driven healthcare systems. As generative AI technologies continue to evolve, future research will likely focus increasingly on establishing robust ethical frameworks, regulatory standards, and evaluation mechanisms capable of supporting the responsible integration of LLMs into clinical practice and medical education.

Limitations and Future Directions

Despite providing a comprehensive overview of the research landscape, this study has several limitations. First, the analysis was restricted to publications indexed in the Web of Science Core Collection and limited to English-language articles and reviews.

Second, bibliometric methods mainly capture patterns of publication output, collaboration structures, and citation relationships, and therefore cannot fully reflect the conceptual depth of ethical arguments presented in individual studies. To mitigate this limitation, the present study combined multiple analytical approaches, including keyword co-occurrence clustering, cross-disciplinary citation flow analysis, and disciplinary diversity indices, which together provide a more comprehensive quantitative representation of the intellectual structure of the field.

Finally, research on large language models in healthcare is evolving rapidly, particularly following the emergence of gen-

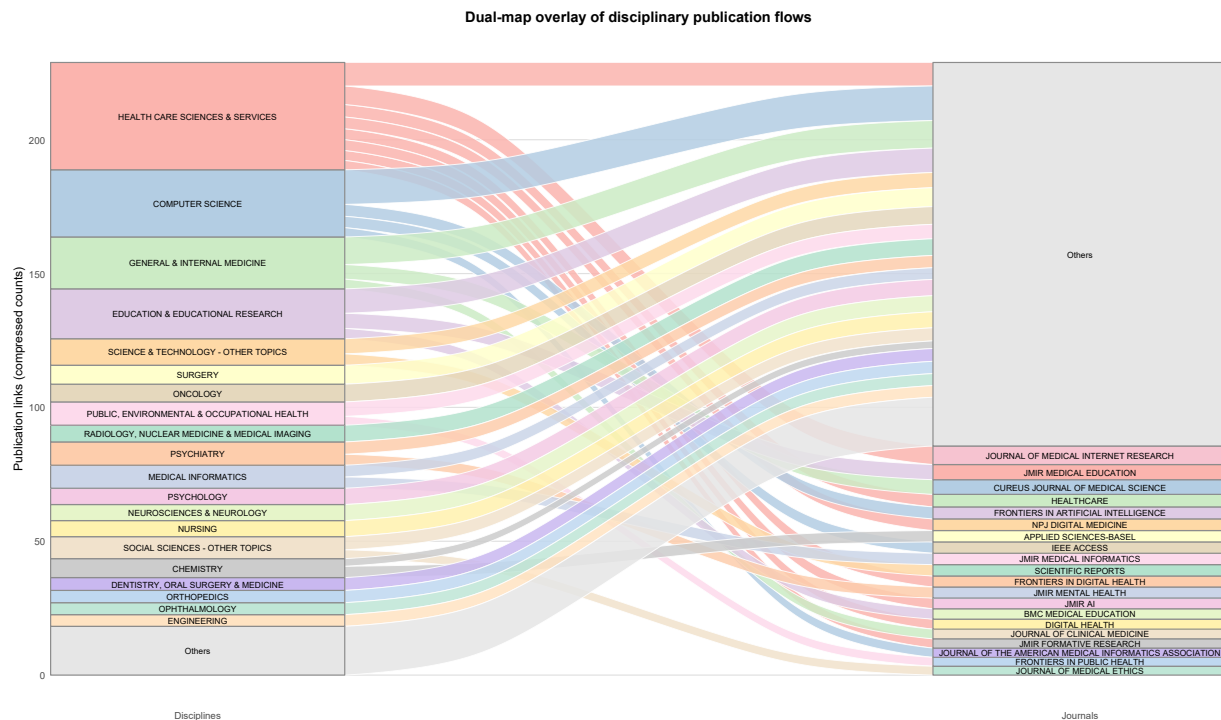


Figure 7. Dual-map overlay of disciplinary publication flows (Discipline → Journal)

The left column displays the primary subject categories (disciplines) of the analyzed publications, while the right column lists the most frequently represented journals. The width of each flow corresponds to the number of publication links between a discipline and a journal (square-root scaled for visualization). To maintain clarity, only the top 20 disciplines and top 20 journals are retained, with remaining items grouped under “Others.” This visualization reveals the distribution of disciplinary outputs across core journals and highlights the degree of concentration versus dispersion in publication venues.

erative AI systems after 2022. As a result, newly published studies may further reshape the thematic structure identified in this analysis. Future research may integrate bibliometric mapping with qualitative or systematic reviews to further examine the development of ethical frameworks and governance strategies for large language models in healthcare.

Declaration

All the authors have declared that there are no conflicts of interest.

REFERENCES

- Achille, A., Kearns, M., Klingenberg, C., & Soatto, S. (2024). A Conceptual Framework for Ethical Issues and their Relationships in the Use of LLMs in Education: Focusing on the Privacy Dimension. *Proceedings of the National Academy of Sciences*, 121(18), e2307304121. <https://doi.org/10.1073/pnas.2307304121>
- Al-kfairy, M., Mustafa, D., Kshetri, N., Insiew, M., & Alfandi, O. (2024a). Ethical Challenges and Solutions of Generative AI: An Interdisciplinary Perspective. *Informatics*, 11(3), 58. <https://doi.org/10.3390/informatics11030058>
- Al-kfairy, M., Mustafa, D., Kshetri, N., Insiew, M., & Alfandi, O. (2024b). Ethical Challenges and Solutions of Generative AI: An Interdisciplinary Perspective. *Informatics*, 11(3), 58. <https://doi.org/10.3390/informatics11030058>
- Allen, J. W., Levy, N., & Wilkinson, D. (2026). Empowering Patient Autonomy: The Role of Large Language Models (LLMs) in Scaffolding Informed Consent in Medical Practice. *Bioethics*, 40(2), 183–193. <https://doi.org/10.1111/bioe.70030>
- Allen, J. W., Schaefer, O., Porsdam Mann, S., Earp, B. D., & Wilkinson, D. (2025). Augmenting research consent: Should large language models (LLMs) be used for informed consent to clinical research? *Research Ethics*, 21(4), 644–670. <https://doi.org/10.1177/17470161241298726>
- Alqahtani, T., Badreldin, H. A., Alrashed, M., Alshaya, A. I., Alghamdi, S. S., Bin Saleh, K., Alowais, S. A., Alshaya, O. A., Rahman, I., Al Yami, M. S., & Albekairy, A. M. (2023). The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy*, 19(8), 1236–1242. <https://doi.org/10.1016/j.sapharm.2023.05.016>
- Androulakis, I. P., Cucurull-Sanchez, L., Kondic, A., Mehta, K., Pichardo, C., Pryor, M., & Renardy, M. (2025). The dawn of a new era: Can machine learning and large lan-

Table 3 Disciplinary diversity index based on cross-disciplinary citation distributions.

Rank	Discipline	Cross-citations	Targets (k)	H	Hnorm
1	DENTISTRY, ORAL SURGERY & MEDICINE	30	4	1.3286	0.9584
2	PSYCHOLOGY	90	5	1.5349	0.9537
3	MEDICAL INFORMATICS	87	5	1.5229	0.9463
4	MATHEMATICAL & COMPUTATIONAL BIOLOGY	49	4	1.3101	0.9451
5	NURSING	53	5	1.4619	0.9083
6	NEUROSCIENCES & NEUROLOGY	77	3	0.9924	0.9033
7	GASTROENTEROLOGY & HEPATOLOGY	60	5	1.4487	0.9001
8	SURGERY	78	4	1.1999	0.8655
9	PSYCHIATRY	74	4	1.1839	0.854
10	ENGINEERING	53	4	1.1122	0.8023
11	CHEMISTRY	69	4	1.1098	0.8006
12	RADIOLOGY, NUCLEAR MEDICINE & MEDICAL IMAGING	52	3	0.8654	0.7877
13	ONCOLOGY	144	5	1.2501	0.7768
14	SOCIAL SCIENCES - OTHER TOPICS	41	3	0.8435	0.7678
15	PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH	68	4	1.0536	0.76
16	EDUCATION & EDUCATIONAL RESEARCH	219	7	1.4435	0.7418
17	HEALTH CARE SCIENCES & SERVICES	558	11	1.7036	0.7104
18	GENERAL & INTERNAL MEDICINE	285	8	1.428	0.6867
19	COMPUTER SCIENCE	476	13	1.7611	0.6866
20	SCIENCE & TECHNOLOGY - OTHER TOPICS	101	5	1.0937	0.6795
21	Others	990	14	1.6611	0.6294

Note: Cross-citations denotes the total number of cross-disciplinary citation links for each discipline, excluding within-discipline self-citations. Targets (k) denotes the number of distinct target disciplines cited. H represents the Shannon diversity index calculated from the distribution of cross-disciplinary citation flows. Hnorm represents the normalized diversity index, calculated as $H_{norm} = \frac{H}{\log k}$, to adjust for differences in the number of target disciplines. Unlike Table 2, diversity indices were computed using the full set of disciplines to avoid truncation bias and to provide a more comprehensive assessment of interdisciplinary breadth. Higher values of H and Hnorm indicate broader citation distributions across disciplines, whereas lower values indicate more concentrated citation orientations.

- guage models reshape QSP modeling? *Journal of Pharmacokinetics and Pharmacodynamics*, 52(4), 36. <https://doi.org/10.1007/s10928-025-09984-5>
- Barr, A. A., Guo, E., Karmur, B. S., & Sezgin, E. (2025). Synthetic neurosurgical data generation with generative adversarial networks and large language models: an investigation on fidelity, utility, and privacy. *Neurosurgical Focus*, 59(1), E17. <https://doi.org/10.3171/2025.4.FOCUS25225>
- Bhayana, R. (2024). Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. *Radiology*, 310(1), e232756. <https://doi.org/10.1148/radiol.232756>
- Boscardin, C. K., Gin, B., Golde, P. B., & Hauer, K. E. (2024). ChatGPT and Generative Artificial Intelligence for Medical Education: Potential Impact and Opportunity. *Academic Medicine*, 99(1), 22–27. <https://doi.org/10.1097/ACM.0000000000005439>
- Caglayan, A., Slusarczyk, W., Rabbani, R. D., Ghose, A., Papadopoulos, V., & Boussios, S. (2024). Large Language Models in Oncology: Revolution or Cause for Concern? *Current Oncology*, 31(4), 1817–1830. <https://doi.org/10.3390/curroncol31040137>
- Cascella, M., Montomoli, J., Bellini, V., & Bignami, E. (2023). Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *Journal of Medical Systems*, 47(1), 33. <https://doi.org/10.1007/s10916-023-01925-4>
- Choudhury, A., & Shamszare, H. (2023). Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis. *Journal of Medical Internet Research*, 25, e47184. <https://doi.org/10.2196/47184>
- Dave, T., Athaluri, S. A., & Singh, S. (2023). ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in Artificial Intelligence*, 6, 1169595. <https://doi.org/10.3389/frai.2023.1169595>
- De Freitas, J., Uğuralp, A. K., Oğuz-Uğuralp, Z., & Puntoni, S. (2024). Chatbots and mental health: Insights into the safety of generative AI. *Journal of Consumer Psychology*, 34(3), 481–491. <https://doi.org/10.1002/jcpy.1393>

- Ferrara, E. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, 6(1), 3. <https://doi.org/10.3390/sci6010003>
- Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *Npj Digital Medicine*, 6(1), 75. <https://doi.org/10.1038/s41746-023-00819-6>
- Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). Large Language Models for Mental Health Applications: Systematic Review. *JMIR Mental Health*, 11, e57400. <https://doi.org/10.2196/57400>
- Jia, J., & Zhao, S. (2025). Ethical and Legal Governance of Generative AI in Chinese Healthcare. *Journal of Multidisciplinary Healthcare*, Volume 18, 5405–5419. <https://doi.org/10.2147/JMDH.S541271>
- Le-Khac, U. N., & Truong, V. N. X. (2025). A survey on large language models unlearning: Taxonomy, evaluations, and future directions. *Artificial Intelligence Review*, 58(12), 399. <https://doi.org/10.1007/s10462-025-11376-7>
- Lin, S., Pandit, S., Tritsch, T., Levy, A., & Shoja, M. M. (2024). What Goes In, Must Come Out: Generative Artificial Intelligence Does Not Present Algorithmic Bias Across Race and Gender in Medical Residency Specialties. *Cureus*. <https://doi.org/10.7759/cureus.54448>
- Liu, Y. (2023). *Summary of ChatGPT-Related research and perspective towards the future of large language models*.
- Lucas, H. C., Upperman, J. S., & Robinson, J. R. (2024). A systematic review of large language models and their implications in medical education. *Medical Education*, 58(11), 1276–1285. <https://doi.org/10.1111/medu.15402>
- Maity, S., & Saikia, M. J. (2025). Large Language Models in Healthcare and Medical Applications: A Review. *Bioengineering*, 12(6), 631. <https://doi.org/10.3390/bioengineering12060631>
- Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *Npj Digital Medicine*, 6(1), 120. <https://doi.org/10.1038/s41746-023-00873-0>
- Moreno, E., Alvarez-Lozada, L. A., Arrambide-Garza, F. J., Quiroga-Garza, A., & Elizondo-Omana, R. E. (2023). Comment on Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* 2023, 11, 887. *Healthcare*, 11(21), 2819. <https://doi.org/10.3390/healthcare11212819>
- Ong, J. C. L., Chang, S. Y.-H., William, W., Butte, A. J., Shah, N. H., Chew, L. S. T., Liu, N., Doshi-Velez, F., Lu, W., Savulescu, J., & Ting, D. S. W. (2024). Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*, 6(6), e428–e432. [https://doi.org/10.1016/S2589-7500\(24\)00061-X](https://doi.org/10.1016/S2589-7500(24)00061-X)
- Raman, R. (2025). Transparency in research: An analysis of ChatGPT usage acknowledgment by authors across disciplines and geographies. *Accountability in Research*, 32(3), 277–298. <https://doi.org/10.1080/08989621.2023.2273377>
- Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare*, 11(6), 887. <https://doi.org/10.3390/healthcare11060887>
- Sandmann, S., Hegselmann, S., Fujarski, M., Bickmann, L., Wild, B., Eils, R., & Varghese, J. (2025). Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nature Medicine*, 31(8), 2546–2549. <https://doi.org/10.1038/s41591-025-03727-2>
- Seo, J., Choi, D., Kim, T., Cha, W. C., Kim, M., Yoo, H., Oh, N., Yi, Y., Lee, K. H., & Choi, E. (2024). Evaluation Framework of Large Language Models in Medical Documentation: Development and Usability Study. *Journal of Medical Internet Research*, 26, e58329. <https://doi.org/10.2196/58329>
- Sung, J. (2023). Artificial intelligence in medicine: Ethical, social and legal perspectives. *Annals of the Academy of Medicine, Singapore*, 52(12), 695–699. <https://doi.org/10.47102/annals-acadmedsg.2023103>
- Tahir, H., Özdemir, M. K., & Alhaji, R. (2025). The role of LLM-powered chatbots in assisting elderly people: Systematic review. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 15(1), 26. <https://doi.org/10.1007/s13721-025-00698-9>
- Temizsoy Korkmaz, F., Ok, F., Karip, B., & Keleş, P. (2025). A structured evaluation of LLM-generated step-by-step instructions in cadaveric brachial plexus dissection. *BMC Medical Education*, 25(1), 903. <https://doi.org/10.1186/s12909-025-07493-0>
- Trujillo, D., Wang, D., Bahr, N., Yi, T., Hsieh, J., Cho, B., Meckler, G., Hansen, M., Eriksson, C., Kim, K. S., Bedrick, S., Jiang, X., & Guise, J.-M. (n.d.). *A medically-grounded LLM agent-based tool to detect patient safety events in medical records*.
- Volkmer, S., Meyer-Lindenberg, A., & Schwarz, E. (2024). Large language models in psychiatry: Opportunities and challenges. *Psychiatry Research*, 339, 116026. <https://doi.org/10.1016/j.psychres.2024.116026>
- Wang, H., Wu, W., Dou, Z., He, L., & Yang, L. (2023). Performance and exploration of ChatGPT in medical examination, records and education in Chinese: Pave the way for medical AI. *International Journal of Medical Informatics*, 177, 105173. <https://doi.org/10.1016/j.ijmedinf.2023.105173>
- Wang, L., Bhanushali, T., Huang, Z., Yang, J., Badami, S., & Hightow-Weidman, L. (2025). Evaluating Generative AI in Mental Health: Systematic Review of Capabilities and Limitations. *JMIR Mental Health*, 12, e70014–e70014. <https://doi.org/10.2196/70014>
- Workum, J. D., Van De Sande, D., Gommers, D., & Van Genderen, M. E. (2025). Bridging the gap: A practical step-

- by-step approach to warrant safe implementation of large language models in healthcare. *Frontiers in Artificial Intelligence*, 8, 1504805. <https://doi.org/10.3389/frai.2025.1504805>
- Wu, J., Wu, X., Zheng, Y., & Yang, J. (2025). Clinical pathway-aware large language models for reliable and transparent medical dialogue. *Journal of Biomedical Informatics*, 172, 104942. <https://doi.org/10.1016/j.jbi.2025.104942>
- Xu, T., Weng, H., Liu, F., Yang, L., Luo, Y., Ding, Z., & Wang, Q. (2024). Current Status of ChatGPT Use in Medical Education: Potentials, Challenges, and Strategies. *Journal of Medical Internet Research*, 26, e57896. <https://doi.org/10.2196/57896>
- Yu, E., Chu, X., Zhang, W., Meng, X., Yang, Y., Ji, X., & Wu, C. (2025). Large Language Models in Medicine: Applications, Challenges, and Future Directions. *International Journal of Medical Sciences*, 22(11), 2792–2801. <https://doi.org/10.7150/ijms.111780>
- Zhui, L., Fenghe, L., Xuehu, W., Qining, F., & Wei, R. (2024). Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education: Viewpoint. *Journal of Medical Internet Research*, 26, e60083. <https://doi.org/10.2196/60083>
- Zohny, H., Allen, J. W., Wilkinson, D., & Savulescu, J. (2025). Which AI doctor would you like to see? Emulating healthcare provider–patient communication models with GPT-4: proof-of-concept and ethical exploration. *Journal of Medical Ethics*, jme-2024-110256. <https://doi.org/10.1136/jme-2024-110256>