

Large Language Models in Healthcare: A Bibliometric Analysis and Examination of Research Trends

Gülcan Gencer¹, Kerem Gencer²

¹Department of Biostatistics and Medical Informatics, Afyonkarahisar Health Sciences University, Faculty of Medicine, Afyonkarahisar, Turkey;

²Department of Computer Engineering, Afyon Kocatepe University, Faculty of Engineering, Afyonkarahisar, Turkey

Correspondence: Gülcan Gencer, Department of Biostatistics and Medical Informatics, Afyonkarahisar Health Sciences University, Faculty of Medicine, Afyonkarahisar, Turkey, Email gencergulcan@gmail.com

Background: The integration of large language models (LLMs) in healthcare has generated significant interest due to their potential to improve diagnostic accuracy, personalization of treatment, and patient care efficiency.

Objective: This study aims to conduct a comprehensive bibliometric analysis to identify current research trends, main themes and future directions regarding applications in the healthcare sector.

Methods: A systematic scan of publications until 08.05.2024 was carried out from an important database such as Web of Science. Using bibliometric tools such as VOSviewer and CiteSpace, we analyzed data covering publication counts, citation analysis, co-authorship, co-occurrence of keywords and thematic development to map the intellectual landscape and collaborative networks in this field.

Results: The analysis included more than 500 articles published between 2021 and 2024. The United States, Germany and the United Kingdom were the top contributors to this field. The study highlights that neural network applications in diagnostic imaging, natural language processing for clinical documentation, and patient data in the field of general internal medicine, radiology, medical informatics, health care services, surgery, oncology, ophthalmology, neurology, orthopedics and psychiatry have seen significant growth in publications over the past two years. Keyword trend analysis revealed emerging sub-themes such as clinical research, artificial intelligence, ChatGPT, education, natural language processing, clinical management, virtual reality, chatbot, indicating a shift towards addressing the broader implications of LLM application in healthcare.

Conclusion: The use of LLM in healthcare is an expanding field with significant academic and clinical interest. This bibliometric analysis not only maps the current state of the research, but also identifies important areas that require further research and development. Continued advances in this field are expected to significantly impact future healthcare applications, with a focus on increasing the accuracy and personalization of patient care through advanced data analytics.

Keywords: large language models, chatbot, healthcare, artificial intelligence, clinical applications, diagnosis, treatment recommendations

Introduction

Large language models (LLMs) are advanced artificial intelligence systems specifically trained using extensive datasets that encompass deep learning and natural language processing (NLP) techniques. These models are particularly recognized for their ability to interpret and generate human language effectively. In recent years, LLMs such as OpenAI's GPT series have demonstrated remarkable capabilities in performing tasks like answering text-based queries, summarizing texts, and translating languages by intricately modeling different language aspects.^{1,2}

In healthcare, LLMs contribute significantly across various domains including medical information retrieval, patient data analysis, and clinical decision support systems. For instance, Esteva et al (2019) demonstrated how LLMs can aid in

diagnosing diseases by processing information from medical literature, illustrating their utility in extracting and interpreting complex medical data. Furthermore, these models are instrumental in developing personalized treatment plans by evaluating extensive patient data, thus enhancing the personalization of healthcare treatments.³

LLMs also play a pivotal role in proposing potential treatment options for specific conditions through the analysis of medical research articles. For example, Roberts et al (2021)⁴ highlighted how LLMs manage complex medical data and integrate into clinical decision-making processes, proving essential in fields such as oncology where they assist in discovering new therapeutic methods.⁵

During the COVID-19 pandemic, LLMs were utilized to model virus spread and evaluate public health interventions, showcasing their capacity to support epidemiological analysis and public health strategy formulation.⁶ The ability of LLMs to streamline operations and reduce costs further promotes their adoption across the healthcare sector, promising significant improvements in efficiency.⁷

The integration of LLMs in healthcare is transformative, with the potential to substantially enhance clinical practices and patient outcomes. These technologies not only improve the delivery of healthcare but also have the capacity to revolutionize the entire healthcare ecosystem.⁸ Over the last decade, the emergence of AI and LLMs at the forefront of health sciences has led to significant advancements, particularly in processing medical texts, defining diseases, and suggesting treatment methodologies. However, the utilization of bibliometric analysis remains crucial for evaluating the spread and impact of these technologies within the healthcare industry. This methodological approach helps in understanding research trends, publication distribution, and academic interactions within the discipline.⁴ It is important to differentiate between general AI or deep learning applications and LLMs. While both are under the AI umbrella, deep learning techniques are broader and encompass various applications beyond language processing, such as analyzing physiological signals in healthcare settings as discussed by Faust et al (2018).⁸

Similarly, Roberts et al (2021) explore the potential of large language models and artificial intelligence to meet health information retrieval and needs during the COVID-19 pandemic.⁴ This study offers methods on how to structure information retrieval systems to enhance the collection and analysis of pandemic-related data.

Jiang and colleagues' (2017) study details the past and present use of AI in healthcare and its future potential, explaining how AI technologies have evolved in fields of medical diagnosis, treatment, and disease management.⁹ Miotto et al (2018) discuss the current status, opportunities, and challenges of deep learning technologies in healthcare.⁷

Kumar and his team's work analyzes the applications of deep learning in healthcare and the technical and ethical challenges associated with these technologies. The primary objective of these studies is to provide an overview of the progress in current research into the use of large language models in healthcare, identifying the main areas of interest and highlighting emerging trends. This analysis aims to assist new researchers in better understanding the potential research opportunities in this area.

This bibliometric study provides a broad perspective on how large language models (LLMs) are being used in the healthcare sector. The increasing research and publication activities over the past three years reflect efforts to harness the potential of artificial intelligence and machine learning technologies to increase diagnostic accuracy, personalize treatment, and improve patient care outcomes. Research has been largely concentrated in the United States, with significant contributions from Europe and Asia, demonstrating the field's global interest and collaborative efforts.

The study found that LLMs extend beyond data processing to direct patient and clinical interactions. Despite this, challenges such as data privacy, ethical concerns, and compliance with healthcare settings remain. Overcoming these challenges is critical to fully realizing the potential of LLMs to transform healthcare.

In conclusion, this analysis has established a solid foundation for future research by delving deeply into the applications of LLMs in the healthcare sector. The successful integration of LLMs requires ongoing research and collaboration, a process that aims to bridge the gap between theoretical models and clinical practices and improve healthcare delivery.

Materials and Methods

Data Sources for Understanding Research Trends of Large Language Models in Healthcare

The primary data sources for this study included Clarivate Analytics' Web of Science (WoS) Core Collection. This multi-disciplinary and authoritative journal citation database was chosen due to its comprehensive coverage of high-impact journals, conferences, and books across various fields. Specifically, the Science Citation Index Expanded (SCIE), Social Science Citation Index (SSCI), Emerging Sources Citation Index (ESCI), and Conference Proceedings Citation Index - Science (CPCI-S) were selected as the main data sources. These indexes were selected because they provide the most up-to-date and broad-reaching literature in the fields of health and technology, making them ideal for understanding emerging research trends and the interactions of large language models (LLMs) in the healthcare sector. The use of WoS was further justified by its robust tools for citation analysis, allowing for a deeper understanding of the impact and evolution of research topics over time. The SCIE, SSCI, ESCI, and CPCI-S indexes were specifically chosen to ensure a comprehensive capture of relevant literature across both established and emerging fields, as well as to include high-impact conference proceedings that often report cutting-edge research.

Data Collection and Search Strategies for Analyzing the Uses of Large Language Models in Medicine

Data collection was conducted in a single session on May 8, 2024, to maintain consistency and avoid the potential impact of data updates during the analysis. The search strategy was meticulously developed to capture the breadth of research on the application of large language models in healthcare. The search query was constructed using terms that reflect the scope of the study, including "NLP", "ChatGPT", and "LLM". These terms were chosen based on their relevance to the study's focus on large language models and their established presence in both academic and industry research. The query returned a total of 2805 records. To ensure the quality and relevance of the analysis, the following criteria were applied: Only articles and reviews published in English were included, to maintain linguistic consistency and accessibility. The inclusion was limited to records from the SCIE, SSCI, ESCI, and CPCI-S indexes, ensuring that only high-quality, peer-reviewed literature was considered. After a manual relevance assessment, which involved careful reading of titles and abstracts, 577 records were deemed relevant. These included 315 original research articles, 89 reviews, and 87 early access and other types of documents. This selection process ensured that the dataset was not only comprehensive but also highly relevant to the study's objectives.

Bibliometric Analysis and Visualization

Basic bibliometric analyses, including the number of publications, citation counts, and global research trends, were initially performed using the built-in tools of the Web of Science platform. For more advanced analyses, including the mapping of academic collaborations and the identification of key research themes, the data was exported to the InCites system, a sophisticated tool for research performance evaluation. The exported data was then analyzed using VOSviewer, a software application specifically designed for constructing and visualizing bibliometric networks. VOSviewer was used to develop various types of bibliometric networks, such as citation networks, co-citation networks, and bibliographic coupling networks. These networks were essential for understanding the relationships between key publications, authors, and research institutions within the field of large language models in healthcare. Additionally, VOSviewer was utilized to create co-occurrence networks of keywords, enabling the identification of emerging trends and themes within the literature. The decision to use VOSviewer was based on its strong capabilities in handling large datasets and its effectiveness in visualizing complex bibliometric relationships, which are crucial for a comprehensive understanding of the field's development.^{10,11}

Citation Metrics and Indexes

Several citation metrics and indexes were calculated to assess the impact and relevance of the selected publications:

Citation Average (CI)

This metric represents the mean number of citations per publication, providing a general measure of impact.

Citation Amount

The total number of citations received by articles in the Web of Science Core Collection, offering a measure of the overall influence of the selected body of work.

International Collaboration

This metric was used to identify research articles authored by contributors from multiple countries, indicating the global nature and collaboration within the field.

h-Index

Developed by J. Hirsch, the h-index was calculated to evaluate the productivity and citation impact of individual researchers within the dataset.

Highest Cited Publications

This index highlights the top 1% of articles in terms of citation count, showcasing the most influential works.

Journal Normalized Citation Rate (JNCI)

JNCI was used to adjust the citation count of an article based on the journal in which it was published, the publication year, and the document type, providing a normalized impact measure.

Subject Normalized Citation Impact (CNCI)

This metric was employed to assess the impact of publications by adjusting citation counts according to subject, year, and document type, offering a standardized comparison across different research areas. JIF: Journal Impact Factor, JCI: Journal Citation Index.

The use of these metrics was crucial in ensuring a thorough and objective evaluation of the literature, allowing for a nuanced understanding of the influence and trends within the field of large language models in healthcare.

Results

Impact of Large Language Models in Health Research and Publication Trends

Referring to the strategic flowchart shown in [Figure 1](#), during the last 4 years (from 2021 to 2024) we have obtained a total of 577 publications in SCIE, SSCI, ESCI and CPCI-S indexes, including 315 original research and 89 review articles. The earliest research on the application of large language models in healthcare is “Evaluation of large language models in natural language processing of PET/CT free-text reports” by Bradshaw and Cho, published in the Journal of Nuclear Medicine in 2021.

This publication received a total of 1 citation. After this date, in 2022, in the journal “Frontiers in Genetics”, Mai et al “TSSNote-CyaPromBERT: Development of an integrated platform for highly accurate promoter prediction and visualization of *Synechococcus* sp. and *Synechocystis* sp. through a state-of-the-art natural language processing model BERT” by Mai et al. In 2023, this number suddenly increased to 337 and in 2024, there were 238 publications in total until May 8, not even half of the year. The number of publications has increased rapidly in the last 2 years ([Figure 2](#)). The number of citations has also been increasing in proportion to the number of publications over the years. According to a specific date, there were 19 publications identified as highly cited by “Clarivate Analytics”. The highest number of citations as of the search date was 101 citations for the article titled “Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers” published in 2023, and all publications were cited a total of 2637 times. The average was 4.57 per citation and the H-index was calculated as 24.

Analysis of Countries or Regions

Countries and regions were analyzed using the Clarivate Incites system. According to [Table 1](#), the countries with more than 1000 publications in total include the United States of America (USA) and the United Kingdom (218 and 54

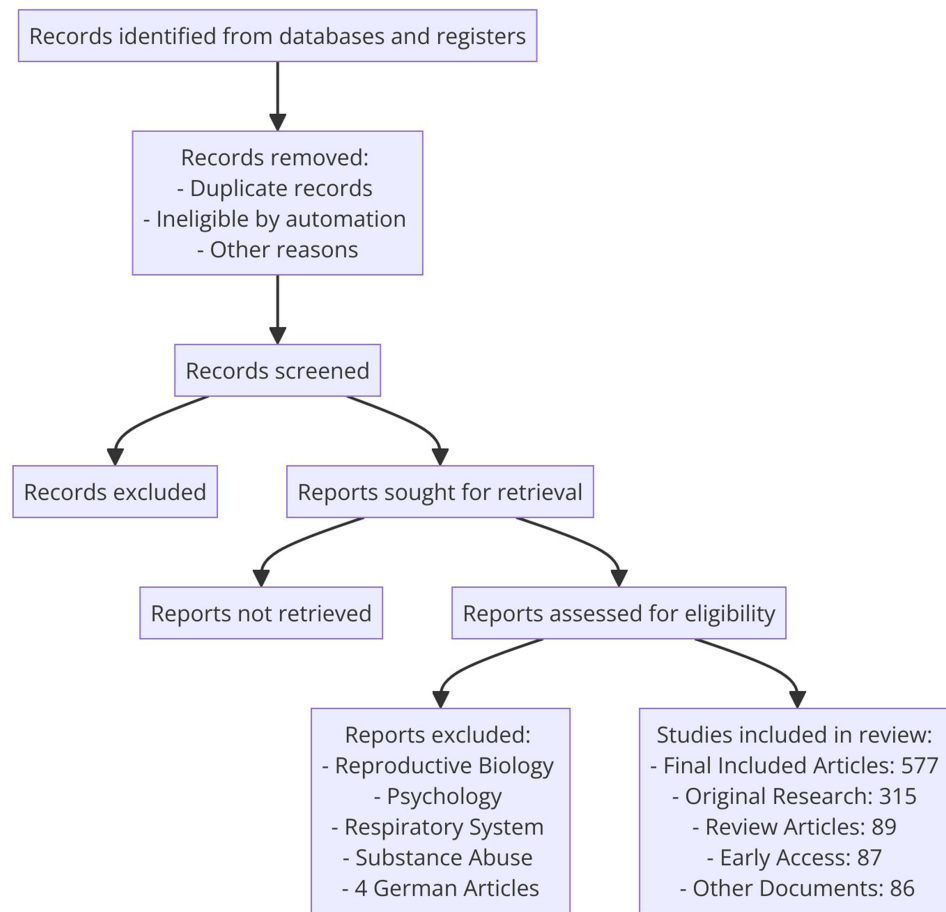


Figure 1 Flowchart for the study.

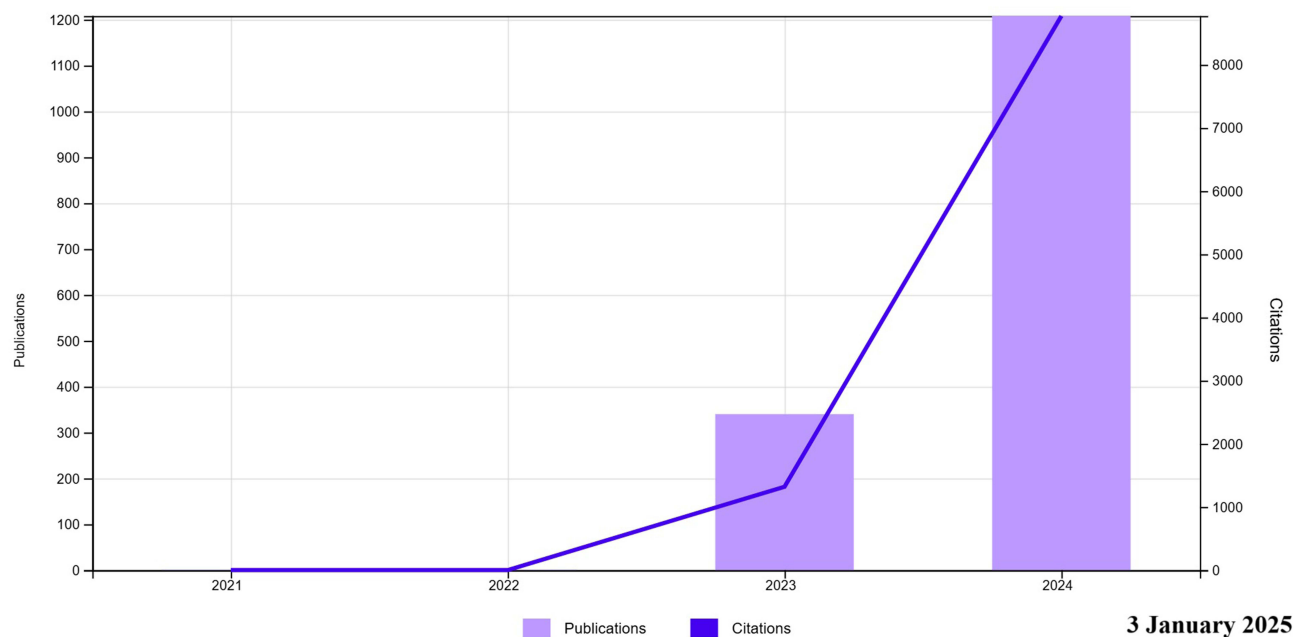


Figure 2 Global trend of publications and citations.

Table 1 Top 10 Effective Countries/Regions Related to Large Language Models in Medicine

Name	Web of Science Documents	Times Cited	Highly Cited Papers	Documents in Top 10%	% Docs Cited	Citation Impact
USA	218	1207 (53.67%)	13 (5.96%)	76(34.86%)	53.67	5.54(8.87)
UNITED KINGDOM	54	185(57.41%)	2 (3.7%)	16(29.63%)	57.41	3.43(6.74)
ENGLAND	53	185(58.49%)	2(3.7%)	16(30.19%)	58.49	3.49(6.87)
GERMANY (FED REP GER)	47	106(55.32%)	0(0%)	13(27.66%)	55.32	2.26(4.07)
ITALY	33	58(33.33%)	0(0%)	9(27.27%)	33.33	1.76(4.62)
INDIA	31	79(45.16%)	1(3.23%)	7(22.58%)	45.16	2.55(3.83)
CANADA	25	73(48%)	1(4%)	6(24%)	48.0	2.92(7.95)
SWITZERLAND	20	143(60%)	1(5%)	9(45%)	60.0	7.15(11.06)
CHINA MAINLAND	19	119(57.89%)	2(10.53%)	6(31.58%)	57.89	6.26(7.34)
AUSTRALIA	18	111(77.78%)	1(5.56%)	12(66.67%)	77.78	6.17(9.25)

respectively). While the USA ranks first in terms of the number of publications, the United Kingdom ranks first in the number of citations with 1207 citations. Among countries with 100 or more citations, the United Kingdom ranks first (185, 185 respectively).

The USA is the leader in this field with 218 Web of Science documents. It is also the most cited country with 1207 citations. Mainland China makes the list with 19 documents and is cited 119 times. This shows a very high citation rate and impact. The US has 13 highly cited papers at 5.96%. This shows that the country is conducting high quality research in this area. The US stands out in this category with 34.86% of its publications. Switzerland has a high percentage in this area with 45% of the 20 documents. Switzerland has the highest value in this area with an average citation impact of 7.15. Among other countries, the USA and Australia also stand out with high citation impact (Table 1). In Figure 3A, the US is shown in dark blue, indicating that it has a high number of documents. Other shades of blue on the map represent the number of Web of Science documents in other countries. For example, regions such as Europe and China also show shades of blue, indicating that these regions have a significant amount of publications. Mainland China is represented by a relatively dark blue on the map, indicating that its scientific activity is increasing. North America and Europe are shown in more intense blue, indicating that science and research activity is more intense in these regions. Africa, the Middle East and some Asian regions are represented in lighter colors, indicating a lower number of documents in these regions. Figure 3B shows the degree of cooperation between countries or regions, conditional on at least 1 publication. Lines between nodes indicate co-authorship between countries, with a thicker line indicating stronger cooperation. The US stands out as the largest node on the map, indicating that it cooperates the most and has the most co-authorship. Germany

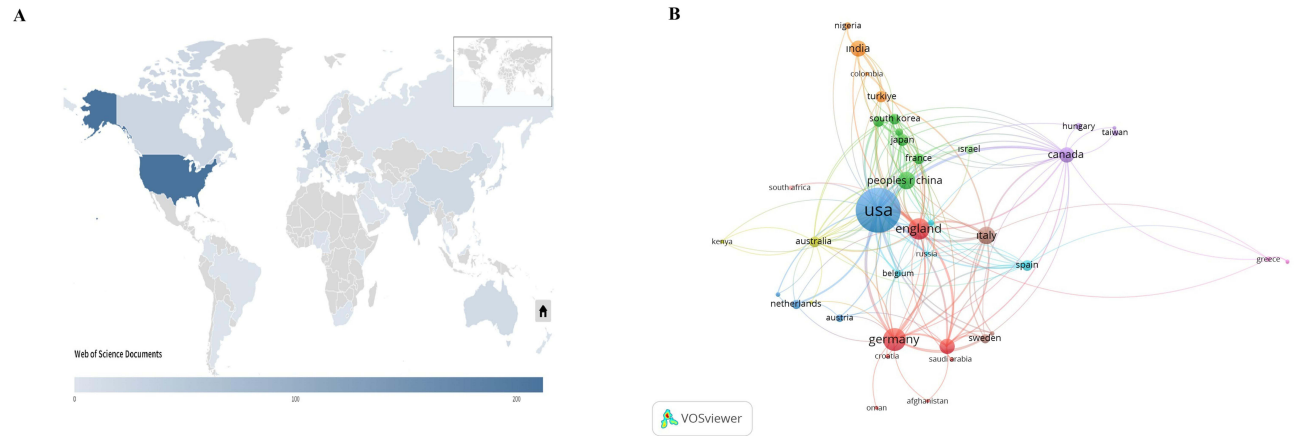


Figure 3 Top countries or regions. (A) Geographic distribution of publications. (B) A network visualization map of countries or regions.

is shown on the map with a prominent red node. There is a dense network between European countries. Similarly, there is also significant cooperation between countries in Asia. Transatlantic links are also evident, with US cooperation with Europe and other regions being quite prominent (Figure 3B).

Analysis of Institutions

Table 2 lists the most productive institutions in the field of medicine related to large language models. The data from this table assesses the scientific impact of these institutions in terms of number of publications, number of citations, citation impact, H-Index values and international collaborations. Harvard University is the most productive institution on this list with 33 papers, which have been cited a total of 131 times. This means that 54.55% of the papers are cited. Stanford University and Harvard Medical School are also highly productive with 26 and 23 papers respectively, with 61.54% and 43.48% of these papers cited. The University of London has the highest citation impact on the list with a citation impact of 4.55 (6.97–5.61), indicating that the quality and impact of its publications is high. Other institutions exhibit values between 1.72 and 3.96 in terms of citation impact, indicating that their impact in their field varies considerably from institution to institution. Harvard University has an H-index of 7, reflecting the sustained impact and quality of research in the field. Other top-ranked institutions, such as Stanford University and Harvard Medical School, also have H-indexes of 5 and 5, indicating that their scientific impact in their field is strong. University College London has the highest percentage in this area, with 66.67% of its credentials in international collaborations. This demonstrates the institution's integration into global science and research networks. Other institutions such as Harvard University and the University of London also have significant international collaborations, with around 33.39% and 60% of their papers, respectively (Table 2).

In Figure 4A, Cornell University and Weill Cornell Medicine have the highest citation shares of 11.5% and 11.4% respectively. New York University has the third highest share at 11.0%. These institutions are notable as institutions with significant influence in the field of medicine related to large language models. Harvard University has a significant citation share of 6.8% in the pie chart. This shows that Harvard's influence in this field is supported not only by the number of publications but also by the citations received. Stanford University is represented by 4.0% in the pie chart. Other institutions in the chart (such as University of Chicago, University of London, Northwestern University) also indicate a certain level of influence. These institutions are represented by a percentage in the range of 4–5% in the pie chart, indicating that they play an important role in the field and are highly cited. Figure 4B shows the collaborations between various universities and research institutions in the field of medicine related to large language models. The nodes in the map represent different institutions, while the lines indicate the presence and intensity of collaborations between them. Harvard Medical School is shown as a prominent and large node on the map, indicating that it plays a central role in collaborations in this field, while other important institutions, such as Stanford University, University of California System, and University College London, which are also featured in Table 2, are also prominent nodes on the map. The lines between these institutions are strong. The map

Table 2 Top 10 Effective Institutions Related to Large Language Models in Medicine

Name	Documents	Times Cited	Citation Impact (CNCI-JNCI)	H-Index	International Collaborations
Harvard University	33	131(54.55%)	3.96(5.84–3.01)	7	13 (39.39%)
Stanford University	26	77(61.54%)	2.96(3.56–1.69)	5	6(23.08%)
Harvard Medical School	23	61(43.48%)	2.65(4.06–2.45)	5	10(43.48%)
University of London	20	91(75%)	4.55(6.97–5.61)	4	12(60%)
University of California System	19	91(52.63%)	3.57(5.63–3.18)	4	4(21.05%)
University College London	15	91(73.33%)	3.66(6.72–4.36)	3	10(66.67%)
Icahn School of Medicine at Mount Sinai	13	91(53.85%)	3.23(5.59–6.27)	4	2(15.38%)
Mayo Clinic	11	91(36.36%)	1.72(2.83–3.51)	2	2(18.18%)
University of Texas System	11	91(54.55%)	2.18(3.72–4.71)	2	5(45.45%)
University of California San Francisco	10	91(40%)	2.9(5.45–4.38)	2	1(10%)

Abbreviations: CNCI, Subject Normalized Citation Impact; JNCI, Journal Normalized Citation Rate.

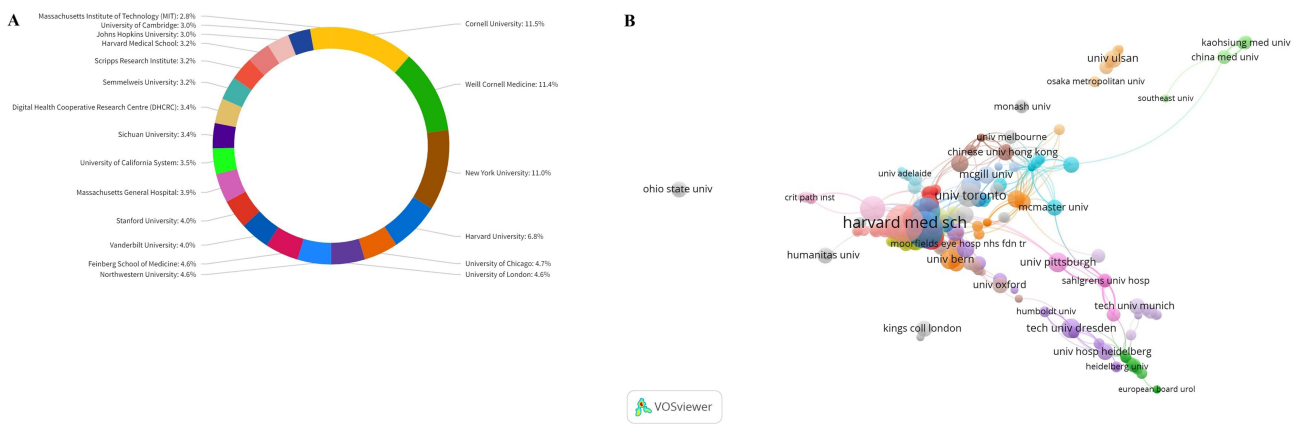


Figure 4 Cooperation and citations between Institutions. **(A)** Most referenced institutions. **(B)** A network visualization map of institutions.

shows complex collaborations between various institutions around the world. For example, there are numerous links between institutions in North America, Europe, Asia and Australia. This shows the prevalence of scientific collaborations on a global scale and that large language patterns are an international endeavor in medical research. In some regions on the map (eg Europe or North America) the lines between nodes are denser. This indicates that regional collaborations are more frequent and perhaps geographical proximity increases the likelihood of collaboration. There also appear to be significant collaborations between technical universities (eg Technical University Munich, Technical University Dresden) and medically focused institutions (eg Moorfields Eye Hospital NHS Foundation Trust). This shows how the disciplines of technology and medicine converge in this area of research.

Analysis of Authors

In Table 3, Seth Ishith is the most prolific author with 8 documents and cited 24 times. Klang Eyal and Sorin Vera also made it to the list with 5 documents respectively, and Sorin Vera in particular achieved a remarkable number of citations with 48 citations. Sorin Vera has an H-index of 3, indicating that he is not only highly cited but also consistent in producing quality publications. Among the other authors, Seth Ishith and Klang Eyal also stand out with an H-index of 3. Sorin Vera stands out with a citation impact of 12.0. With a CNCI of 14.82, this high impact indicates that his work in the field is highly valued by the scientific community. Klang Eyal’s CNCI value is 11.86 and citation impact is 9.6, again showing a high scientific impact. JNCI measures the relative citation success of an author’s publications relative to other works in related fields. Seth Ishith has a very high value of 12.91 on this scale, indicating that his work is more highly cited than other work in the field. Yi Paul H. also stands out with a JNCI value of 10.95.

Table 3 Top 10 Effective Authors Related to Large Language Models in Medicine

Name	Web of Science Documents	Times Cited	H-Index	Citation Impact	CNCI	JNCI
Seth Ishith	8	24	3	3.0	(6.29)	(12.91)
Rozen Warren Matthew	6	16		2.7	(6.31)	(0)
Wiwanitkit V.	6	0	0	0.0	(0)	(0)
Yi Paul H.	5	7	2	1.4	(7.06)	(10.95)
Klang Eyal	5	48	3	9.6	(11.86)	(5.43)
Ross Richard J.	4	4	1	1.0	(2.97)	(12.47)
Venerito Vincenzo	4	14	2	3.5	(4.44)	(4.16)
Sorin Vera	4	48	3	12.0	(14.82)	(6.79)
Kleebayoon Amnuay	4	0	0	0.0	(0)	(0)
Cevik Jevan	3	3	1	1.0	(1.42)	(1.71)

Abbreviations: CNCI, Subject Normalized Citation Impact; JNCI, Journal Normalized Citation Rate.

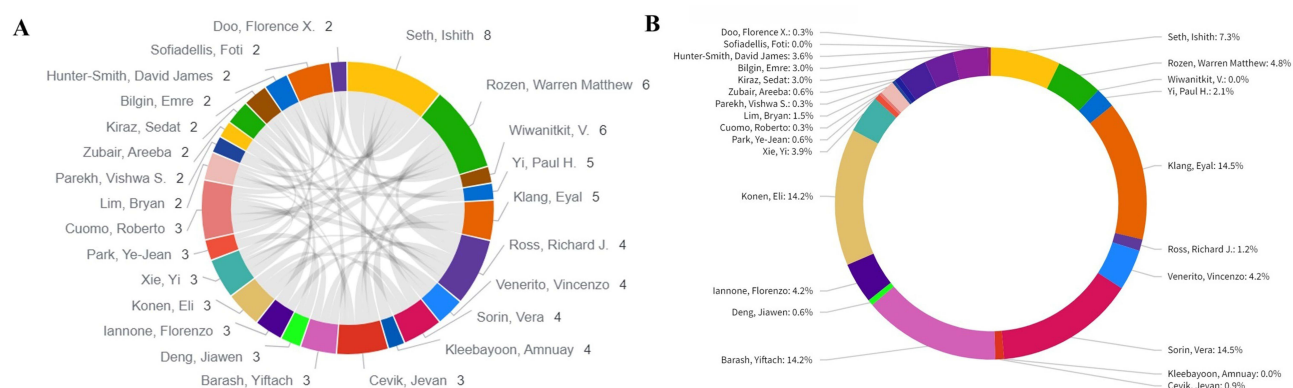


Figure 5 Authors collaborate. (A) The most productive authors. (B) The most cited authors.

In **Figure 5A**, Seth Ishith has the largest slice of the image, with 8 Web of Science documents. The lines in the figure show his collaborations with other authors. These collaborations reflect the interactions with other researchers and the multidisciplinary nature of Ishith's work in this field. Other authors such as Rozen Warren Matthew and Wiwanitkit V. also have significant slices in the image. Both authors have 6 documents, indicating effective and active collaborations. The lines between the various authors in the image indicate the intensity and nature of their collaborations. For example, authors such as Klang Eyal and Sorin Vera have multiple collaborative links with others, indicating that they play important roles as central figures in this field. The density of the lines indicates the frequency of collaborations and the strength of the ties between researchers. Dense lines represent more frequent and possibly deeper collaborations. The image shows a highly integrated community of researchers working on large language models and medicine. Such collaborations can accelerate scientific progress by promoting the exchange of knowledge and the development of innovative ideas. Even authors with fewer publications (eg Cevik Jevan et al) are represented in the image, indicating their small but important contributions.

In **Figure 5B**, Klang Eyal and Sorin Vera have the largest slices of the graph, each with 14.5%. Barash Yiftach has the third largest slice at 14.2%. This indicates that his work in the field is also having a high impact. Seth Ishith has a remarkable share of the pie chart at 7.3%. Rozen Warren Matthew occupies a significant slice of the chart with 4.8%, indicating his citation impact and recognition in the sector. Authors such as Wiwanitkit V. and Kleebyaon Amnuay have very small or no citations at all.

This perhaps indicates that their work is addressed to a narrower audience or that they are newer.

In **Figure 6A**, two distinct groups of cooperation stand out. These groups are shown in different colors (red and green), and the dense network of lines within each group indicates close cooperation between group members. Both groups have central figures within them. For example, in the red group, names such as “Tham Yih Chung” and “Lee,

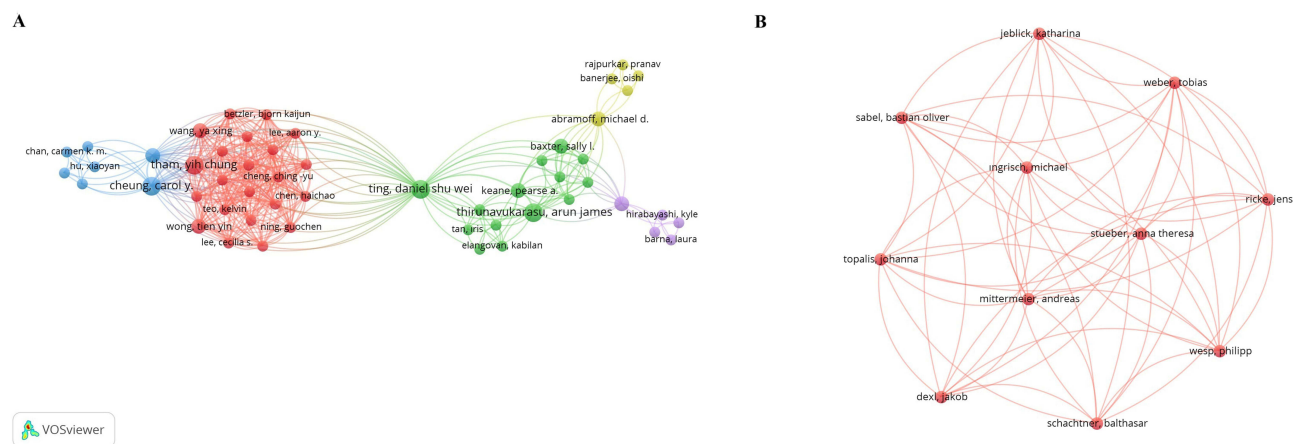


Figure 6 Collaboration among authors. (A) Authors' visualization map. (B) A visual map of the most cited authors.

Aaron Y.”; in the green group, names such as “Ting, Daniel Shu Wei” and “Keane, Pearse A.” play central roles. The lines between the two groups are less dense but exist, indicating that there are also collaborations between these groups. These lines represent integration and knowledge exchange between different research groups or disciplines. Researchers’ names and collaboration structures are likely to be from different countries and the most cited authors are shown in Figure 6B.

Analysis of Journals

In Table 4, the journals that contributed the most to the literature are mainly; “Cureus Journal Of Medical Science”, “Journal Of Medical Internet Research” and “Radiology” journals (26, 15 and 9 respectively) (Table 4).

In Figure 7, the Radiology, Nuclear Medicine and Medical Imaging category has the highest number of citations with 464 citations. The Medical Informatics category comes second with 419 citations. This indicates that large language models are increasingly being used in the processing and management of health information. The Health Sciences and Services and Surgery categories received 381 and 181 citations respectively. The number of citations in these categories indicates that large language models are effective in diagnosing diseases, improving treatment methods and planning surgical procedures. The use of large language models is also notable in specialized fields such as Ophthalmology (eye diseases) and Oncology (cancer research). These fields received 159 and 117 citations respectively, while fields such as Orthopedics, Dermatology and Clinical Neurology received fewer citations. This may indicate that large language models have not yet been widely adopted in these specialized fields, or perhaps their use in these fields has not yet been fully explored.

Figure 8 visualizes the relationships between keywords and research categories used in the medical field related to large language models. In the map, keywords and categories are shown by representing their interconnections with lines and groups with different colors. The red group includes terms such as “clinical management”, “education”, “nlp” (natural language processing), and “artificial intelligence”. The green group includes terms such as “clinical research”, “medicine”, “machine learning”, “chatbots”, and “scientific writing”. This group shows how large language models are used in research, data analysis and scientific communication. There are connections between both groups, bridged by

Table 4 Top 10 Effective Journals Related to Large Language Models in Medicine

Name	Documents	Times Cited	Citation Impact	CNCI	JNCI	Eigen factor	5 Year Impact Factor	Journal Impact Factor	JCI Quartile	JIF Quartile
Cureus Journal Of Medical Science	26	109	4.19	4.44	15.09	0.05	1.2	1.2	Q3	n/a
Journal Of Medical Internet Research	15	146	9.73	11.21	7.33	0.06	7.6	7.4	Q1	Q1
Radiology	9	261	29.00	57.76	13.97	0.05	14.1	19.7	Q1	Q1
Jama-Journal Of The American Medical Association	7	57	8.14	10.51	3.85	0.30	81.4	120.7	Q1	Q1
Npj Digital Medical	7	177	25.29	27.21	7.48	0.02	15.2	15.2	Q1	Q1
Korean Journal Of Radiology	6	26	4.33	8.88	4.70	0.01	4.7	4.8	Q1	Q1
Journal Of Cancer Research And Clinical Oncology	5	29	5.80	7.49	3.76	0.01	3.8	3.6	Q2	Q2
Lancet Digital Health	5	52	10.40	4.17	1.21	0.01	30.9	30.8	Q1	Q1
Aesthetic Surgery Journal	4	24	6.00	11.17	3.41	0.01	3.5	3.1	Q1	Q1
European Radiology	4	23	5.75	6.71	6.12	0.04	5.5	5.9	Q1	Q1

Abbreviations: CNCI, Subject Normalized Citation Impact; JNCI, Journal Normalized Citation Rate; JIF, Journal Impact Factor; JCI, Journal Citation Indicator.

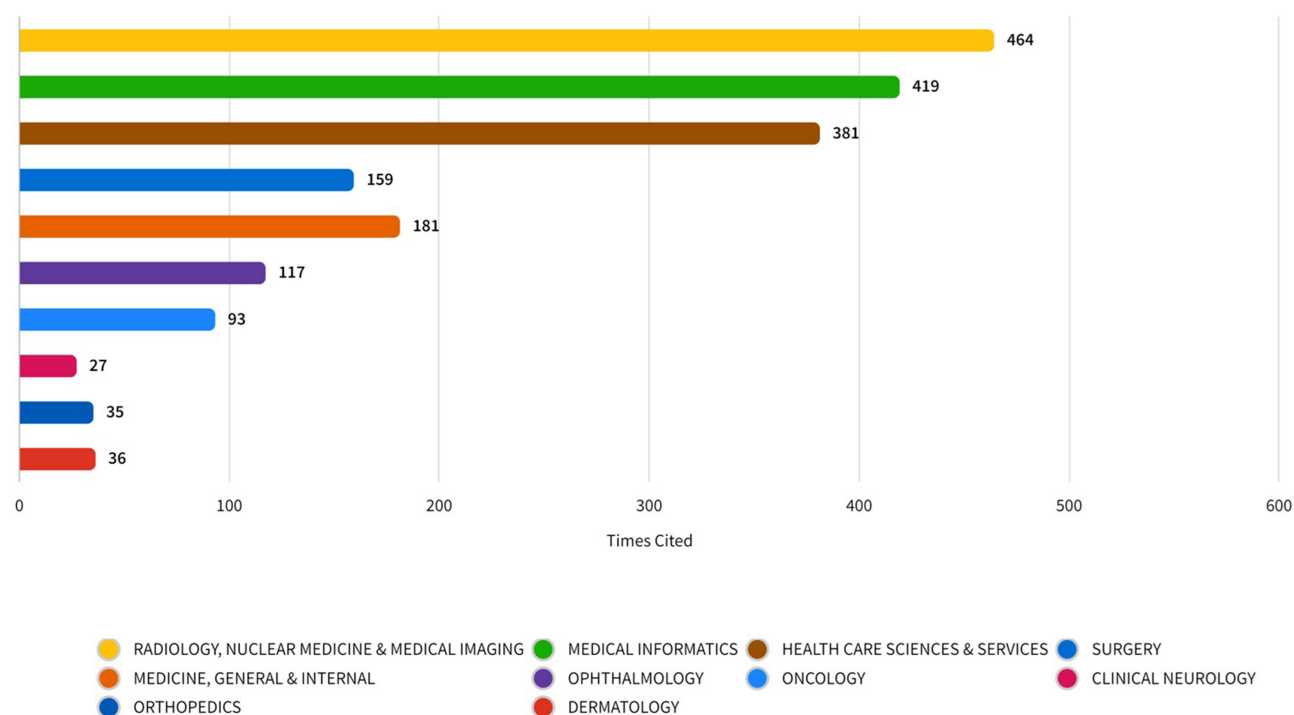


Figure 7 Web of Science research areas.

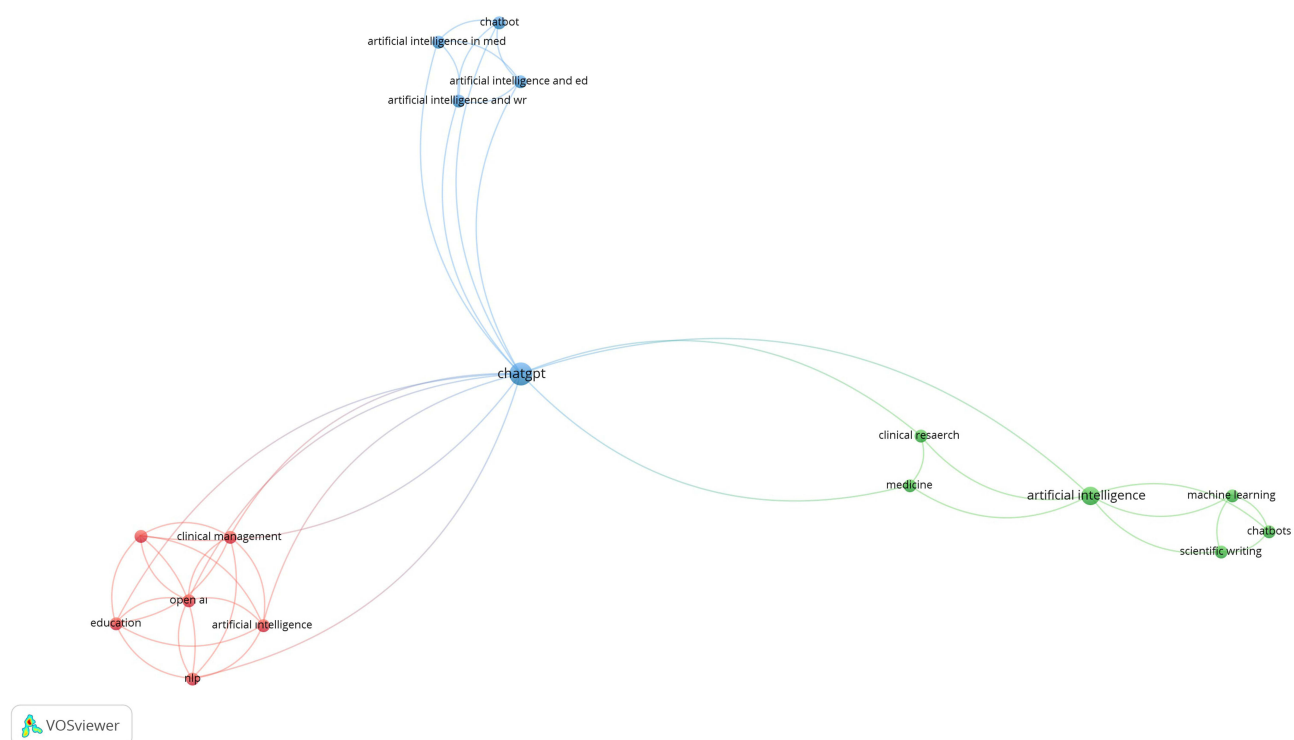


Figure 8 Keyword co-occurrence analysis and network visualization map.

common terms such as “artificial intelligence” and “machine learning”. Terms such as “chatbot” and “ChatGPT” are keywords that link clinical management and education groups with clinical research and medicine groups. This shows how chatbots and automated text generation tools are used in medical applications. The visual highlights how large language models are integrated in an interdisciplinary approach in medicine. For example, the combination of the terms

“clinical research” and “machine learning” indicates the importance of these technologies in data analysis. The co-occurrence of the terms “scientific writing” and “chatbots” indicates that the use of AI-based tools in scientific content production is on the rise.

Discussion

Large Language Models (LLMs) are increasingly being employed in healthcare to enhance clinical practices, diagnostic accuracy, and treatment methodologies. Since the late 1990s, advancements in artificial intelligence and natural language processing (NLP) technologies have significantly improved the utilization of health information. Over the past decade, the integration of LLMs into healthcare has grown exponentially, reflecting their transformative potential in this domain. The surge in research focusing on LLM applications in health sciences is evident, with the majority of studies originating from prominent research centers in the United States, Europe, and Asia. These models have demonstrated their utility in diverse areas, including extracting meaningful insights from clinical notes, querying data from medical knowledge bases, and managing patient records. Hosseini et al (2024)¹² demonstrated the application of LLMs in clinical abbreviation disambiguation, while Cascella et al (2024)¹³ explored their development and utility in medical applications over a one-year period.

Wang et al (2024)¹⁴ highlighted the ability of LLMs to structure free-text surgical records, enhancing decision-making in stroke management. Their study emphasized the potential of these models to improve outcomes by efficiently processing unstructured data, a crucial challenge in healthcare. Park et al (2024)¹⁵ comprehensively assessed the landscape of LLM research and their use in clinical settings, highlighting their impact on healthcare. The study by Nerella et al (2024)¹⁶ detailed how transformer architectures underpin LLMs in healthcare information processing. Similarly, Denecke et al (2024)¹⁷ showcased LLM-driven innovations in diagnosis and treatment, emphasizing their potential in data analysis. Yan et al (2024)¹⁸ and Benítez et al (2024)¹⁹ examined how LLMs support phenotyping algorithms and medical education, respectively. Additionally, Preiksaitis et al (2024)²⁰ highlighted LLM applications in emergency medicine, while Raja et al (2024)²¹ evaluated their utility in categorizing and analyzing ophthalmology-related scientific literature. Similarly, Gencer and Gencer (2024)²² explored the transformative potential of AI in medical education by comparing ChatGPT’s performance with medical faculty graduates in specialization exams, highlighting its role in enhancing training methodologies and knowledge assessment. Such studies underscore the wide-ranging impact of LLMs on patient outcomes, particularly in enhancing diagnostic accuracy and personalizing treatments. Furthermore, a review by Nazi and Peng (2024)²³ synthesized current LLM applications in medical imaging and diagnostics.

Their findings aligned with Singh et al (2023),²⁴ demonstrating enhanced diagnostic precision in oncology through AI-driven web applications. These insights reveal LLMs’ capacity to integrate diverse data sources, supporting complex decision-making processes in clinical environments.

LLMs also present unprecedented opportunities for administrative efficiency, as demonstrated by Gebreab et al (2024),²⁵ who explored LLM-based frameworks for automating healthcare tasks. Meskó and Topol (2023)²⁶ emphasized the importance of regulatory oversight in ensuring the ethical use of AI in healthcare, underscoring the need for robust frameworks to mitigate risks associated with data security and algorithmic bias. The transformative impact of LLMs is evident in specialized fields. Rajkomar et al (2019)²⁷ detailed how AI improves disease diagnosis through medical data analysis. Gawehn et al (2016)²⁸ explored their role in decoding the molecular basis of diseases, enabling personalized treatments. These contributions are complemented by recent findings from Singh et al (2023),²⁴ who demonstrated enhanced diagnostic precision in oncology using AI-driven web applications. Despite these advances, challenges persist. The high computational demands of training and deploying LLMs, coupled with ethical concerns related to patient data privacy, highlight areas requiring further research. The successful integration of LLMs into healthcare demands translational research that bridges the gap between theoretical innovations and practical applications. For instance, Mann et al (2020)¹ highlighted the need for domain-specific training to enhance model accuracy, while Radford et al (2019)² demonstrated the versatility of unsupervised LLMs in diverse contexts. Going forward, interdisciplinary collaborations between technologists, clinicians, and policymakers will be essential. Recent works, such as those by Huang et al (2020)⁵ and Omiye et al (2024),²⁹ emphasize the role of AI in improving diagnostic workflows and reducing healthcare disparities. These efforts must be complemented by targeted investments in education and training to build workforce

capacity in AI adoption. This discussion highlights the multifaceted potential of LLMs in healthcare while acknowledging the challenges that lie ahead. By fostering a collaborative ecosystem, future research can unlock the full potential of LLMs, revolutionizing healthcare delivery worldwide.

General Information

Large Language Models (LLMs) represent a significant advancement in the field of artificial intelligence, particularly in natural language processing (NLP). These models, such as GPT-3, BERT, and their successors, are designed to understand, generate, and process human language in a way that closely mimics human communication. The development and application of LLMs have revolutionized various sectors, including healthcare, by providing innovative solutions for data analysis, information retrieval, and decision support.^{12,15} Transformers, introduced in the paper “Attention Is All You Need” by Vaswani et al, were a milestone in NLP. Transformers use a self-attention mechanism that allows the model to weigh the relative importance of different words in a sentence, improving contextual understanding of the language.³⁰ This architecture laid the foundation for the development of powerful LLMs such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-Trained Transformers). LLMs are characterized by their large scale, often containing billions of parameters. This scale allows them to store large amounts of linguistic information and capture complex patterns in data. LLMs typically undergo a two-stage training process. During pre-training, the model is exposed to large text corpora to learn general language patterns. Fine-tuning involves further training on specific datasets to tailor the model to specific tasks or domains, such as medical text analysis. Unlike traditional models, LLMs can understand the context of words within a sentence, making them capable of more accurate language understanding and production. This is crucial for applications in areas such as healthcare, where context can significantly alter the meaning of information. LLMs are extremely versatile and can be applied to a wide range of tasks, including text production, translation, summarization, and question answering. This versatility makes them valuable tools in a variety of applications. LLMs have shown significant potential in healthcare, improving several aspects of medical practice and research: LLMs can help create and organize clinical notes, ensuring that documentation is comprehensive and accurately reflects patient encounters. These models can extract relevant information from unstructured texts, such as electronic health records (EHRs), to support clinical decision-making and research. LLMs can empower chatbots and virtual assistants by providing accurate information and support to patients in real time. By analyzing large datasets, LLMs can help predict disease outbreaks, patient outcomes, and potential complications, enabling proactive healthcare management. Large Language Models represent a transformative technology in healthcare that offers innovative solutions to improve clinical practice, enhance patient care, and advance medical research.^{18,27}

Key Challenges in the Current State of AI in Healthcare

The integration of artificial intelligence (AI) into healthcare presents significant potential to transform clinical practice, yet it also brings a host of challenges that must be addressed to realize this potential fully. One of the primary challenges lies in the clinical integration of complex AI tools, such as polygenic risk scores and deep learning models, which require sophisticated interpretation and communication strategies to be effectively utilized by healthcare providers and understood by patients. Additionally, the lack of diversity in training datasets limits the generalizability of AI models across different populations, necessitating efforts to include more diverse genetic and demographic data. The fusion of multi-modal data, interoperability, and standardization of AI systems remain technical hurdles that impede widespread adoption in clinical settings. Moreover, the black-box nature of many AI models, particularly deep learning and transformer-based models, raises concerns about transparency and interpretability, which are crucial for gaining the trust of clinicians and patients alike. Regulatory and ethical considerations also pose significant barriers, with the need for clear guidelines to address data privacy, algorithmic bias, and accountability in AI-driven decisions. As AI tools increasingly rely on large-scale, high-quality data, issues of data availability, quality, and the substantial computational resources required for advanced models like transformers further complicate their deployment. Ultimately, addressing these challenges—ranging from technical and computational issues to ethical and regulatory concerns—will be critical for advancing AI in healthcare and ensuring that these technologies can be safely and effectively integrated into routine clinical practice.

Limitations

A systematic scan of publications until 08.05.2024 was carried out using the Web of Science database. While comprehensive, this database may not include all relevant publications from other major sources like PubMed or Scopus, potentially omitting valuable biomedical research. Additionally, restricting the language to English may have excluded significant studies in other languages, biasing results toward English-speaking regions.

The manual screening process, though necessary, may have introduced selection bias, particularly where subjective judgment influenced inclusion decisions. Furthermore, reliance on tools like VOSviewer for mapping research landscapes heavily depends on the quality of input data, potentially overlooking nuanced trends or less-cited emerging studies.

This study primarily uses publication and citation metrics to provide a descriptive overview of LLM applications, which may not directly reflect the practical or clinical effectiveness of these technologies. Metrics such as citation counts and the H-index often favor older publications, potentially underestimating the transformative potential of newer research.

Lastly, the ambiguity in the definition of “large language models” complicates the interpretation of findings, particularly regarding their intended use in research versus clinical applications. Only a limited number of AI-driven medical devices, approximately 950, have received FDA approval as of 2024, highlighting the gap between experimental research and clinical implementation. Future studies should integrate data from multiple databases, include non-English research, and distinguish between experimental and clinically validated applications to address these gaps effectively.

Conclusion

This bibliometric analysis provided a comprehensive overview of the research landscape on the application of large language models (LLMs) in healthcare. The study highlighted a significant increase in LLM-related research activities over the last three years, particularly in fields such as radiology, internal medicine, and oncology. The findings reveal a global interest, with notable contributions from the United States, Europe, and Asia, but also underscore the need for greater international collaboration to leverage diverse expertise.

Emerging sub-themes, such as the use of LLMs in clinical management, medical education, and patient data analysis, demonstrate the expanding scope of their applications beyond traditional data processing. Despite these advancements, challenges related to data privacy, ethical concerns, and model robustness remain key obstacles to their widespread adoption in healthcare settings.

In summary, while LLMs hold transformative potential for healthcare, their successful integration requires continued research, development, and collaboration. Future efforts should prioritize translational research that bridges theoretical models and practical, clinically applicable solutions to improve patient outcomes and healthcare delivery.

Data Sharing Statement

The dataset used and/or analyzed during this study is available upon reasonable request from the author of the article.

Ethics Approval

The datasets utilized in this study were obtained from a publicly accessible database. Importantly, our research did not include any experimental procedures involving human or animal subjects. Therefore, ethics board approval was not necessary.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work. All authors have read and agreed to the published version of the manuscript.

Funding

This study did not receive any external financial support.

Disclosure

The authors state that there are no personal relationships or financial interests that could seem to have influenced the work reported in this paper.

References

- Mann B, Ryder N, Subbiah M, et al. Language models are few-shot learners. *arXiv preprint arXiv:200514165*. 2020
- Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI blog*. 2019;1(8):9.
- Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24–29. doi:10.1038/s41591-018-0316-z
- Roberts K, Alam T, Bedrick S, et al. Searching for scientific evidence in a pandemic: an overview of TREC-COVID. *J Biomed Inform*. 2021;121:103865. doi:10.1016/j.jbi.2021.103865
- Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges. *Cancer Lett*. 2020;471:61–71. doi:10.1016/j.canlet.2019.12.007
- Abd-Alrazaq A, Alajlani M, Alhuwail D, et al. Artificial intelligence in the fight against COVID-19: scoping review. *J Med Int Res*. 2020;22(12):e20756. doi:10.2196/20756
- Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Briefings Bioinf*. 2018;19(6):1236–1246. doi:10.1093/bib/bbx044
- Faust O, Hagiwara Y, Hong TJ, Lih OS, Acharya UR. Deep learning for healthcare applications based on physiological signals: a review. *Comput Methods Programs Biomed*. 2018;161:1–13. doi:10.1016/j.cmpb.2018.04.005
- Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vascular Neurol*. 2017;2(4):230–243. doi:10.1136/svn-2017-000101
- Tang M, Mu F, Cui C, et al. Research frontiers and trends in the application of artificial intelligence to sepsis: a bibliometric analysis. *Front Med*. 2023;9:1043589. doi:10.3389/fmed.2022.1043589
- Van Eck NJ, Waltman L. VOSviewer manual. *Manua VOSviewer Version*. 2011;1:1.
- Hosseini M, Hosseini M, Javidan R. Leveraging Large Language Models for Clinical Abbreviation Disambiguation. *J Med Syst*. 2024;48(1):27. doi:10.1007/s10916-024-02049-z
- Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. *J Med Syst*. 2024;48(1):22. doi:10.1007/s10916-024-02045-3
- Wang M, Wei J, Zeng Y, et al. Precision Structuring of Free-Text Surgical Record for Enhanced Stroke Management: a Comparative Evaluation of Large Language Models. *J Multidiscip Healthc*. 2024;Volume 17:5163–5175. doi:10.2147/JMDH.S486449
- Park Y-J, Pillai A, Deng J, et al. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Med Inf Decis Making*. 2024;24(1):72. doi:10.1186/s12911-024-02459-6
- Nerella S, Bandyopadhyay S, Zhang J, et al. Transformers and large language models in healthcare: a review. *Artif Intell Med*. 2024;2024:102900.
- Denecke K, May R, Rivera Romero O. LLMHealthGroup. Potential of Large Language Models in Health Care: delphi Study. *J Med Int Res*. 2024;26:e52399. doi:10.2196/52399
- Yan C, Ong HH, Grabowska ME, et al. Large language models facilitate the generation of electronic health record phenotyping algorithms. *J Am Med Inf Assoc*. 2024;2024:ocae072.
- Benítez TM, Xu Y, Boudreau JD, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. *J Am Med Inf Assoc*. 2024;31(3):776–783. doi:10.1093/jamia/ocad252
- Preiksaitis C, Ashenburg N, Bunney G, et al. The Role of Large Language Models in Transforming Emergency Medicine: scoping Review. *JMIR Medical Inform*. 2024;12:e53787. doi:10.2196/53787
- Raja H, Munawar A, Mylonas N, et al. Automated Category and Trend Analysis of Scientific Articles on Ophthalmology Using Large Language Models: development and Usability Study. *JMIR Format Res*. 2024;8(1):e52462. doi:10.2196/52462
- Gencer G, Gencer K. A comparative analysis of ChatGPT and medical faculty graduates in medical specialization exams: uncovering the potential of artificial intelligence in medical education. *Cureus*. 2024;16(8):e66517. doi:10.7759/cureus.66517
- Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. *MDPI*. 2024;2024:57.
- Singh A, Randive S, Breggia A, Ahmad B, Christman R, Amal S. Enhancing Prostate Cancer Diagnosis with a Novel Artificial Intelligence-Based Web Application: synergizing Deep Learning Models, Multimodal Data, and Insights from Usability Study with Pathologists. *Cancers*. 2023;15(23):5659. doi:10.3390/cancers15235659
- Gebreab SA, Salah K, Jayaraman R, Ur Rehman MH, Ellaham S. LLM-Based Framework for Administrative Task Automation in Healthcare. *IEEE*. 2024;2024:1–7.
- Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *Npj Digital Med*. 2023;6(1):120. doi:10.1038/s41746-023-00873-0
- Rajkumar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347–1358. doi:10.1056/NEJMr1814259
- Gawehn E, Hiss JA, Schneider G. Deep learning in drug discovery. *Mol Inform*. 2016;35(1):3–14. doi:10.1002/minf.201501008
- Omiye JA, Ghanzouri I, Lopez I, et al. Clinical use of polygenic risk scores for detection of peripheral artery disease and cardiovascular events. *PLoS One*. 2024;19(5):e0303610. doi:10.1371/journal.pone.0303610
- Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inform Process Syst*. 2017;30:5998–6008.

Journal of Multidisciplinary Healthcare

Dovepress
Taylor & Francis Group

Publish your work in this journal

The Journal of Multidisciplinary Healthcare is an international, peer-reviewed open-access journal that aims to represent and publish research in healthcare areas delivered by practitioners of different disciplines. This includes studies and reviews conducted by multidisciplinary teams as well as research which evaluates the results or conduct of such teams or healthcare processes in general. The journal covers a very wide range of areas and welcomes submissions from practitioners at all levels, from all over the world. The manuscript management system is completely online and includes a very quick and fair peer-review system. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <https://www.dovepress.com/journal-of-multidisciplinary-healthcare-journal>