Scientometrica 1 (2025) 165–174

Contents lists available at GrowingScience

Scientometrica

homepage: www.GrowingScience.com/sci

A scientometric review of highly cited literature on large language models in healthcare: Trends, applications, and intellectual structure

Dmaithan Almajali[a*]

[a]*Management Information System Department, Faculty of business, Applied Science private University (ASU) Amman, Jordan*

C H R O N I C L E

Article history:
Received: November 5, 2025
Received in revised format: November 5, 2025
Accepted: November 31, 2025
Available online: November 31, 2025

Keywords:
Large Language Models
Healthcare
Scientometrics
Artificial Intelligence
ChatGPT
Clinical Applications
Medical Ethics
Systematic Review
Research Trends

A B S T R A C T

The introduction of Large Language Models (LLMs) in the healthcare sector is a major step, which can for sure transform all aspects of medicine including practice, research, and education. This review of publications presents the first 200 most cited articles which were obtained from a Scopus search on the terms "large language models" and "healthcare". The purpose was to elucidate the knowledge and trends in this fast-growing area. The analysis looks at the yearly publications, the main journals, the most prominent authors and institutions, the leading research areas, the methodological approaches used, and the ethical and regulatory issues that are the most talked about. The results show an increase in the number of scholars interested in this field, especially in 2023 and 2024. Besides that, it was found that there is a lot of high-impact publications going on in the leading multidisciplinary and specialized medical journals areas. The largest research areas that were found are: clinical trials and their outcomes, ethical and governance frameworks, educational integration, and technological advancements and surveys. The research is mainly focused on testing the performance of LLMs like ChatGPT in particular medical tasks but in the background, there are profound concerns about their accuracy, bias, and safety. This review presents the current knowledge state, points out the most active and leading research fronts, and recognizes the gaps with future directions hence it offers the most basic reference for researchers, doctors, and policymakers who will be dealing with the LLMs incorporation into the healthcare system.

© 2025 by the authors; licensee Growing Science, Canada.

1. Introduction

The development of huge language models (LLMs), which are a type of AI system that is trained using extensive text corpora, has brought along a revolutionary change in different industries. One of the most impacted, and necessarily so, is the medical field where LLMs could be utilized in various areas such as reading and understanding clinical notes, reviewing scientific papers, and even communicating with patients (Thirunavukarasu et al., 2023). These models are expected to provide support in making clinical decisions, reduce the time taken for administrative tasks, offer tailored patient education, and promote faster medical research (Yu et al., 2023; Zhang & Boulos, 2023).

The rate at which research is being conducted in this area has reached an astonishing level. After the introduction of ChatGPT (Generative Pre-trained Transformer) and similar models to the public, the researchers have participated continuously in a highly dynamic cycle of exploration, assessment, and critique. The outcome has been a large and quickly growing collection of publications. Nevertheless, the very rapidity and quantity of this production can make it hard for one to find their way around. It is most critical for researchers, practitioners, and investors to be aware of the main ideas, the most important works, and the intellectual framework of the domain in order to spot knowledge gaps, prevent overlapping efforts, and give an accurate direction to the ensuing research.

* Corresponding author.
E-mail address: almjalidmaithan@yahoo.com (D. Almajali)

ISSN 3115-8455 (Online) - ISSN 3115-8447 (Print)
© 2025 by the authors; licensee Growing Science, Canada
doi: 10.5267/j.sci.2025.5.003

Quantitative analysis of scientific communication and literature, which is the study of Scientometrics, allows very efficiently to map and analyze the emerging research areas. Application of the techniques of Scientometrics, one can change the way the narrative synthesis is done and provide a data-driven overview of the trends in publications, networking of collaborations, and the evolution of a field's concepts.

The present paper offers a scientometric assessment of the very influential literature in the area between LLMs and healthcare. The paper's main goal is to reveal the existing mental framework of this domain by examining a carefully selected collection of 200 highly cited papers. In particular, this review has the following objectives:

1. Quantify the temporal and spatial distribution of publications.
2. Identify the most influential journals, authors, and documents.
3. Map the dominant research themes and application areas.
4. Analyze the methodological approaches and key challenges highlighted in the literature.
5. Discuss the ethical, regulatory, and practical considerations that are shaping the field.

This review, which offers a systematic and quantitative summary of the most influential research, not only sets a standard for comprehension of the current situation of LLMs in healthcare but also directs future scientific investigations and practical implementations.

2. Methodology

2.1 Data Source and Search Strategy

This scientometric analysis is based on data obtained from the Scopus database, which is considered to be one of the largest and most trusted abstract and citation databases containing peer-reviewed literature. The search was executed with the keywords "large language models" AND "healthcare". Through this search method, a wide but still pertinent set of publications was identified that lie at the heart of the interdisciplinary field. The first search brought in more than 2000 records, which is indicative of the great amount of scholarly attention directed toward the issue.

2.2 Data Filtering and Selection

In order to analyze the most important and influential works, the dataset was narrowed down to just the first 200 highly cited references from the results. The term "highly cited" was concretely defined by using Scopus's internal citation metrics during the search, which ranks the records by the number of citations. This method guarantees the corpora being analyzed are of the publications that have drawn the most researchers' attention and gained their influence, thus giving access to the core debates and the key papers in the discipline. Along with the data, a file with the bibliographical metadata of the selected 200 articles was also provided.

2.3 Data Analysis

Analysis was mostly descriptive and taxonomic, focusing on quantitative metrics and thematic categorizations. Consequently, I followed the procedures given below:

- **Bibliometric Analysis:** The years of publication, source titles (journals), authors' names as well as their institutional affiliations were extracted and arranged in a table to detect the trends and significant contributors.
- **Thematic Analysis:** The titles, abstracts, author keywords, and, if the case demanded, the full text of the articles were all scanned in order to recognize and group the main research themes, application areas, and types of studies (for example, review, original research, and editorial). This was a compound process of reading and coding which resulted in the formation of a proper classification scheme that truly represents the literature.
- **Content Analysis:** A qualitative synthesis was carried out to encapsulate the major discoveries, difficulties, and future trends as pointed out in the most influential papers of each thematic cluster by citation count.

It should be emphasized that the current review excludes methods for network analysis, neither visualization maps (e.g., co-citation networks, keyword co-occurrence maps) were produced according to the specifications. The data presented here are only in the form of tables and narrative synthesis.

2.4 Limitations

There are some limitations in this study. One of the shortcomings is the reliance on a single database for the analysis, Scopus, which, while considered comprehensive, may still exclude some relevant publications. Another limitation is the selection of the top 200 highly cited articles, which may lead to a bias towards older, foundational studies as well as those published in high-impact journals, possibly overlooking newer, niche, or applied studies that have not yet gained high citation counts due to being too new or small scale. The "early mover advantage" in a rapidly developing field might distort results. Lastly, the nature of this field is ever-changing, which means that the landscape is in a constant state of flux; thus, this analysis only gives a view of what was like at the time of the dataset.

3. Results

3.1 Publication Trends and Growth

The scrutiny of the publication years regarding the 200 most cited papers shows a strong and sudden burst of interest, which is clearly focused on the years 2023 and 2024. This coincides with the public launch and extensive use of modern LLMs such as ChatGPT (GPT-3.5 and GPT-4) in the last part of 2022 and throughout 2023.

Table 1
Distribution of Highly Cited Publications by Year

Publication Year	Number of Publications	Percentage of Total (%)
2023	92	46
2024	85	42.5
2025	23	11.5
Total	**200**	**100**

Table 1 displays that the year 2023 is by far the most productive year in terms of major publications since it yielded a total of 92 (46.0%) highly cited papers. Year 2024 came next with 85 articles (42.5%) suggesting that the research area had still been very much alive and impactful with a very high level of output. In addition, the appearance of 23 recorded papers from 2025 (11.5%), in a dataset collected at this point in time, is a sign of the early or pre-print issuing of the most promising research, hence, the high-speed nature of the field is further confirmed.

3.2 Core Journals and Publication Venues

The 200 articles published in total came from a varied range of 87 different journals. Nevertheless, the distribution is quite uneven, with a few high-impact journals mainly hosting this research.

Table 2
Top 10 Journals Publishing Highly Cited LLM-in-Healthcare Research

Rank	Journal Title	Number of Publications	Percentage of Total (%)
1	npj Digital Medicine	16	8
2	Nature Medicine	9	4.5
3	Journal of Medical Systems	8	4
4	BMC Medical Informatics and Decision Making	6	3
5	Communications Medicine	4	2
6	Healthcare (Switzerland)	4	2
7	Journal of the American Medical Informatics Association	4	2
8	PLoS Digital Health	4	2
9	Artificial Intelligence Review	3	1.5
10	Diagnostics	3	1.5
	All Other Journals (77)	*139*	*69.5*

Table 2 clearly shows that the Nature Portfolio journals (npj Digital Medicine and Nature Medicine) are the top players as they collectively cover 12.5% of the most-cited literature. Such a situation not only underscores the preference for but also the impact of publications in high-ranking, wide-ranging medical and digital health journals. The general trend of the "All Other Journals" category reflects the interdisciplinary character of the field, revealing that considerable amounts of research were made public via not only clinical specialty journals but also computer science and ethics forums.

3.3 Document Types and Research Focus

The dataset presents a number of document types representing the maturity and nature of the discourse.

Table 3
Distribution of Document Types

Document Type	Number of Publications	Percentage of Total (%)
Article	132	66
Review	42	21
Conference Paper	11	5.5
Editorial	6	3
Short Survey	5	2.5
Note	4	2

The most part (66.0%) is made up of original research articles, which is typical of a young and empirically driven domain where new proofs are quickly produced. The high percentage of reviews (21.0%) represents an effort to concurrently synthesize, summarize and comprehend the fast-paced advances. The inclusion of conference papers, editorials and notes indicates a vibrant and ongoing dialogue that embraces not only the early findings but also the opinions and controversies. A more in-depth analysis of the research focus reveals a number of different categories:

Table 4
Primary Research Focus of the Publications

Research Focus Category	Number of Publications	Percentage of Total (%)	Description
Clinical Application & Evaluation	85	42.5	Studies evaluating LLM performance in diagnosis, treatment planning, clinical reasoning, etc.
Review & Survey	52	26	Systematic, scoping, and narrative reviews; surveys of the field.
Ethics, Governance & Challenges	31	15.5	Papers focused on bias, fairness, privacy, regulation, and workforce impact.
Technical & Methodological Development	20	10	Proposing new models, frameworks, or techniques (e.g., RAG, fine-tuning).
Education & Patient Interaction	12	6	Use of LLMs in medical education, training, and patient communication.

The prominence of "Clinical Application & Evaluation" (42.5%) indicates in the strongest way possible the main research direction of the testing of LLMs' real-world application and safety in the clinical setting. Then, the application of "Review & Survey" (26.0%) is in next position as the community tries to nourish the knowledge together. The area of "Ethics, Governance & Challenges" absorbs and attracts (15.5%), reflecting the high concern over the risks and the necessity of practicing ethics in a considerate manner.

3.4 Influential Authors and Collaborative Networks

Although a complete co-authorship network analysis is not included in this report, the identification of the most prolific authors in the dataset reveals the major contributors. Authoring has become a collective effort in this field as the number of papers with numerous authors, usually coming from different institutions and interdisciplinary teams, is quite high.

Table 5
Authors with Three or More Appearances in the Dataset

Author Name	Number of Publications	Author Name	Number of Publications
Kather, J.N.	6	Daneshjou, R.	3
Bressem, K.K.	4	Giuffrè, M.	3
Celi, L.A.	4	Kresevic, S.	3
Shung, D.L.	4	McCoy, L.G.	3
Adams, L.C.	3	Ting, D.S.W.	3
Busch, F.	3	Truhn, D.	3
Cascella, M.	3		

It is common to see authors such as Jakob N. Kather, Leo A. Celi, and Daniel L. Shung appearing numerous times, mostly by working together in the areas of clinical medicine, oncology, gastroenterology, and AI research. This list is a reflection of the core scientists who are very active and significantly influential in pushing the field forward.

4. Thematic Analysis of the Intellectual Landscape

The examination of the content of the 200 articles made it possible to spot four principal and interrelated thematic groups that outline the present intellectual framework of the LLM investigation in healthcare.

4.1 Cluster 1: Clinical Applications and Task-Specific Evaluations

This is the biggest and the most varied group of research works, which contains the studies that testing the performance of LLMs by means of empirical methods mainly ChatGPT and GPT-4—on particular medical tasks. The assessments are often comparative, with LLM outputs being benchmarked against established guidelines or human expert performance.

- **Medical Knowledge and Reasoning:** A considerable amount of research was dedicated to the evaluation of LLMs on the medical exam questions. In a head-to-head cross-sectional study by Thirunavukarasu et al. (2023), it was shown that LLMs could be at the level of or even surpass expert-level clinical knowledge and reasoning in ophthalmology, with such a conclusion also drawn in other areas. In the same manner, Guillen-Grima et al. (2023) assessed ChatGPT's effectiveness on the Spanish MIR residency exam, revealing encouraging outcomes.

- **Diagnostic Support and Clinical Decision-Making:** This sub-theme research is focused on the use of LLMs in medical diagnosis and treatment recommendations. The studies move from general medical models for disease diagnosis (Liu et al., 2025) to assessments in particular situations like the emergency department (Williams et al., 2024), oncology (Carl et al., 2024), and cardiology (Kusunose et al., 2023). It has been widely accepted that LLMs, although very powerful, are still not fit for human-free clinical application since their recommendations are sometimes incorrect and/or not safe (Au Yeung et al., 2023).

- **Administrative and Documentation Workflow:** One more important area of application is to reduce the load of administrative work through automation. It entails the production of clinical notes (Goyal et al., 2024), the preparation of discharge summaries (Rosenberg et al., 2024), the replying to patient messages (Liu et al., 2024), and the improving of prior authorization processes (Tripathi et al., 2024). These applications can frequently be regarded as the lowest-risk entry points for LLM integration, with a very high promise of efficiency gains.

- **Patient Triage and Communication:** Numerous researches investigated the deployment of LLMs as primary communicative means with patients, like in outpatient reception (Wan et al., 2024) or by answering general medical queries (Lahat et al., 2023). The concern is mainly about precision, compassion, and guidance to the right treatment for patients.

4.2 Cluster 2: Ethical, Equity, and Regulatory Considerations

This cluster tackles the deployment of LLMs in the high-stakes healthcare environment with its associated critical challenges and risks. It is marked by a warning note and demands strict control.

- **Bias and Health Equity:** A remarkable study conducted by Omiye and co-workers (2023) showed that LLMs might introduce race-based medicine, thus maintaining and spreading dangerous stereotypes and falsehoods. Moreover, this research pointed out the possibility of these models being a source of bias, as they are often trained on biased data. Later on, the research community has been scale up their efforts with the aim to develop methods that will help to make such inequity harms visible (Pfohl et al., 2024).

- **Hallucinations and Accuracy:** The inclination of large language models (LLMs) to "hallucinate" or create believable yet erroneous information is a significant issue regarding patient safety (Maleki et al., 2024). Consequently, it has led to investigations in decreasing the adverse effects, one of which is Retrieval-Augmented Generation (RAG), which associates the model's output with verified external knowledge sources (Kresevic et al., 2024; Wang et al., 2024).

- **Privacy and Data Security:** The proprietary LLMs' utilization sparks the concerns regarding the patient data confidentiality that is already mentioned by Das et al. in 2025. On the other hand, this scenario has also stimulated the interest towards the creation of open-source (Labrak et al., 2024) and locally deployable models, apart from the contemplation of the privacy-preserving techniques such as federated learning (Rauniyar et al., 2024).

- **Regulation and Governance:** The development of LLMs has progressed at such a fast pace that the current regulatory frameworks are unable to cope with the situation. Some of the papers that have the most impact have raised the question of regulatory oversight as a necessity (Meskó & Topol, 2023) and at the same time have started to explore the regulations that are just beginning to come about, like the EU AI Act (Busch et al., 2024). Reporting guidelines such as TRIPOD-LLM (Gallifant et al., 2025) and MI-CLEAR-LLM (Park et al., 2024) have been developed, which thus constitute a significant step towards the standardization and quality enhancement of research in this domain.

4.3 Cluster 3: Medical Education and Patient Empowerment

This cluster delves into the possible transformations that LLMs could bring about as facilitators of learning and knowledge sharing among medical professionals and patients.

- **Enhanced Learning for Students and Clinicians:** Largely in part, the LLMs are being recognized as great and useful allies in medical education, wherein a few of their functions are to generate clinical stories (vignettes),

simplify hard subjects, and allow for simulated patient interactions (Yigci et al., 2025; Lower et al., 2023). On the other hand, there have been arguments over the issues of academic dishonesty and dependency on the technology (Cain et al., 2023).

- **Patient Education and Health Literacy:** Research has looked into quality and readability of medical information that LLMs are giving to patients for various conditions from myopia (Biswas et al., 2023) to inflammatory bowel disease (Gravina et al., 2024). The purpose is to evaluate whether they can serve as an additional source of information that would raise patient's health literacy and facilitate the process of making decisions together (Aydin et al., 2024).

4.4 Cluster 4: Technological Foundations, Surveys, and Future Directions

The papers in this cluster strive to lay a foundation for technical knowledge with high-level hedging and various points of view.

- **Comprehensive Surveys and Reviews:** A multitude of influential reviews have holistically presented the state of the art, putting down through the applications, breakthroughs, and difficulties (Thirunavukarasu et al., 2023; Sallam, 2023; Wang & Zhang, 2024; Meng et al., 2024; Nazi & Peng, 2024). These contributions are primary access routes to the subject for the unskilled ones.

- **Technical Advancements:** To mention some examples, the research under this umbrella covers fine-tuning general-purpose LLMs to medical specialties (Liu et al., 2024), inventive architectures, and enhancement of reasoning by means of Chain-of-Thought (Miao et al., 2024; Tang et al., 2024) improvisation among others.

- **Vision for the Future:** The future scenario has been illustrated by some powerful articles, predicting the change from tools to AI agents (Qiu et al., 2024; Mehandru et al., 2024) and the rise of a new paradigm of "large AIl models" or "foundation models" that combine language, vision, and multimodal data for a more comprehensive understanding of health (Qiu et al., 2023; Clusmann et al., 2023).

5. Discussion

This particular scientometric analysis, which is based on 200 highly cited publications, depicts the field as one going through an explosive, dynamic, and a bit cautious phase of growth. The major concentration of research influence in the years 2023-2024 is a direct result of the challenging and at the same time disruptive introduction of large language models (LLMs) that are publicly accessible and quite sophisticated. The conversation is being transformed by the conflicting situation in which the models' major capacity to improve healthcare is demonstrated and their profound limitations and risks are seriously acknowledged.

5.1 Synthesis of Major Trends

This particular scientometric analysis, which is based on 200 highly cited publications, depicts the field as one going through an explosive, dynamic, and a bit cautious phase of growth. The major concentration of research influence in the years 2023-2024 is a direct result of the challenging and at the same time disruptive introduction of large language models (LLMs) that are publicly accessible and quite sophisticated. The conversation is being transformed by the conflicting situation in which the models' major capacity to improve healthcare is demonstrated and their profound limitations and risks are seriously acknowledged.

The second most significant trend is the importance of ethics and governance. Similar to the latter, the incorporation of LLMs into medicine has been constantly and openly scrutinized regarding its ethical and moral aspects right from the start. The references of papers on bias (Omiye et al., 2023), regulation (Meskó & Topol, 2023), and hallucinations are so high that they suggest the scientific community considers tackling these problems as a condition for adoption, thus, not as an afterthought.

5.2 Predominant Challenges and Recurring Themes

Literature generally echoes the problems of competitiveness:

1. **The Accuracy-Safety Paradox:** Nonetheless, large language models (LLMs) are capable of sometimes performing very impressively, but still their failure modes remain unpredictable and the models might be dangerously confident. Thus, the direct reliance on such models for autonomous clinical decision-making is, at this point, not acceptable.

2. **The Data Biasis:** AI models raised at that mass would inherit some social prejudices and then increase these, thus posing a serious threat to health equity.

3. **The Black Box Problem:** The reason for LLM being a black box poses a problem for trust and accountability and it also makes it hard for the doctors to be aware of the model's suggestion "why" at least in this case.

4. **The Implementation Gap:** The difference between proving efficacy in a controlled research environment and safely and effectively integrating a tool into a complex clinical workflow is quite large.

The idea to position the LLM as an "augmenting" tool and not as a substitute for medical staff has been a strong and frequent story (Thirunavukarasu, 2023; Harrer, 2023). The eventual picture that comes out from the literature is one where human-AI cooperation is taking place, wherein LLMs take care of searching for information, summarizing, and generating drafts, thus allowing doctors to concentrate on advanced reasoning, showing compassion, and making difficult choices.

5.3 Gaps and Future Research Directions

According to the analysis, different gaps and future directions have been determined:

- **Longitudinal and Real-World Studies:** It is highly necessary to conduct research that will assess the impacts of LLM integration on patients' long-term outcomes, physician's exhaustion, and the healthcare expenses in actual situations, not just in isolated tasks.

- **Multimodal Integration:** Future developments in healthcare will involve models that are capable of dealing with and reasoning on multimodal data sources like clinical text, medical images, genomic data, and continuous sensor data, thus giving a broader view of the patient (Yildirim et al., 2024).

- **Specialized and Open-Source Models:** The dependence on proprietary models such as GPT-4 brings up issues of reproducibility, expense, and control. Among the most important areas of development in the future will be the creation and assessment of high-quality open-source medical LLMs (for instance, BioMistral by Labrak et al., 2024).

- **Advanced Evaluation Metrics:** It is indispensable to establish uniform measures that not only include accuracy but also take into account clinical reasoning, resistance to adversarial prompts, and fairness among different patient groups.

- **Interdisciplinary Collaboration:** The dataset's most influential piece comes from the close collaboration of computer scientists, clinicians, ethicists, and social scientists. The nurturing of these collaborations will be pivotal in overcoming the complex and various challenges of the future.

6. Conclusion

This scientometric review presents a detailed diagram of the most cited works on applying large language models to the medical field. A distinctive feature of the sector is its spectacular advancement, it being mostly published in the most influential journals, and the balancing act of revealing transformative potential while rigorously interrogating the risks of the new methods. The core of the intellectual structure consists of the four areas of clinical evaluation, ethical consideration, educational innovation, and technological survey. The results interpretation puts LLMs in the position of the probably most important technological shift that the healthcare research community has taken very active and critical engagement with. The way suggested by this key literature is the one of cautious, evidence-based, and ethically grounded integration, not of blind adoption. The power of LLMs use in producing a future of healthcare that is not only more efficient but also more equitable, empathetic, and effective is the ultimate goal common to these 200 papers. Future research that will be based on this foundation will conduct powerful real-world trials, create more sophisticated and specialized models, and the necessary regulatory and ethical frameworks that will ensure this powerful technology gets to fulfil its promise of benefiting all humanity.

References

Au Yeung, J., Kraljevic, Z., Luintel, A., Balston, A., Idowu, E., Dobson, R. J., & Teo, J. T. (2023). AI chatbots not yet ready for clinical use. *Frontiers in Digital Health, 5*, 1161098.

Aydin, S., Karabacak, M., Vlachos, V., & Margetis, K. (2024). Large language models in patient education: a scoping review of applications in medicine. *Frontiers in Medicine, 11*, 1477898.

Biswas, S., Logan, N. S., Davies, L. N., Sheppard, A. L., & Wolffsohn, J. S. (2023). Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. *Ophthalmic and Physiological Optics, 43*(6), 1562–1570.

Busch, F., Kather, J. N., Johner, C., Moser, M., Truhn, D., Adams, L. C., & Bressem, K. K. (2024). Navigating the European Union Artificial Intelligence Act for Healthcare. *npj Digital Medicine, 7*(1), 210.

Cain, J., Malcom, D. R., & Aungst, T. D. (2023). The Role of Artificial Intelligence in the Future of Pharmacy Education. *American Journal of Pharmaceutical Education, 87*(10), 100135.

Carl, N., Schramm, F., Haggenmüller, S., Kather, J. N., Hetz, M. J., Wies, C., Michel, M. S., Wessels, F., & Brinker, T. J. (2024). Large language model use in clinical oncology. *npj Precision Oncology, 8*(1), 240.

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J.-N., Laleh, N. G., Löffler, C. M. L., Schwarzkopf, S.-C., Unger, M., Veldhuizen, G. P., Wagner, S. J., & Kather, J. N. (2023). The future landscape of large language models in medicine. *Communications Medicine, 3*(1), 141.

Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and Privacy Challenges of Large Language Models: A Survey. *ACM Computing Surveys, 57*(6), 152.

Gallifant, J., Afshar, M., Ameen, S., Aphinyanaphongs, Y., Chen, S., Cacciamani, G., Demner-Fushman, D., Dligach, D., Daneshjou, R., Fernandes, C., Hansen, L. H., Landman, A., Lehmann, L., McCoy, L. G., Miller, T., Moreno, A., Munch, N., Restrepo, D., Savova, G., … Celi, L. A. (2025). The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine, 31*(1), 60–69.

Goyal, S., Rastogi, E., Rajagopal, S. P., Yuan, D., Zhao, F., Chintagunta, J., Naik, G., & Ward, J. (2024). HealAI: A Healthcare LLM for Effective Medical Documentation. In *WSDM 2024 - Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (pp. 1167–1168).

Gravina, A. G., Pellegrino, R., Cipullo, M., Palladino, G., Imperio, G., Ventura, A., Auletta, S., Ciamarra, P., & Federico, A. (2024). May ChatGPT be a tool producing medical information for common inflammatory bowel disease patients' questions? An evidence-controlled analysis. *World Journal of Gastroenterology, 30*(1), 17–33.

Guillen-Grima, F., Guillen-Aguinaga, S., Guillen-Aguinaga, L., Alas-Brun, R., Onambele, L., Ortega, W., Montejo, R., Aguinaga-Ontoso, E., Barach, P., & Aguinaga-Ontoso, I. (2023). Evaluating the Efficacy of ChatGPT in Navigating the Spanish Medical Residency Entrance Examination (MIR): Promising Horizons for AI in Clinical Medicine. *Clinics and Practice, 13*(6), 1460–1487.

Harrer, S. (2023). Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine, 90*, 104512.

Kresevic, S., Giuffrè, M., Ajcevic, M., Accardo, A., Crocè, L. S., & Shung, D. L. (2024). Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *npj Digital Medicine, 7*(1), 102.

Kusunose, K., Kashima, S., & Sata, M. (2023). Evaluation of the Accuracy of ChatGPT in Answering Clinical Questions on the Japanese Society of Hypertension Guidelines. *Circulation Journal, 87*(7), 1030–1033.

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., & Dufour, R. (2024). BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (pp. 5848–5864).

Lahat, A., Shachar, E., Avidan, B., Glicksberg, B., & Klang, E. (2023). Evaluating the Utility of a Large Language Model in Answering Common Patients' Gastrointestinal Health-Related Questions: Are We There Yet? *Diagnostics, 13*(11), 1950.

Liu, S., Mccoy, A. B., Wright, A. P., Carew, B., Genkins, J. Z., Huang, S. S., Peterson, J. F., Steitz, B., & Wright, A. (2024). Leveraging large language models for generating responses to patient messages - a subjective analysis. *Journal of the American Medical Informatics Association, 31*(6), 1367–1379.

Liu, X., Liu, H., Yang, G., Jiang, Z., Cui, S., Zhang, Z., Wang, H., Tao, L., Sun, Y., Song, Z., Hong, T., Yang, J., Gao, T., Zhang, J., Li, X., Zhang, J., Sang, Y., Yang, Z., Xue, K., … Wang, G. (2025). A generalist medical language model for disease diagnosis assistance. *Nature Medicine, 31*(3), 932–942.

Lower, K., Seth, I., Lim, B., & Seth, N. (2023). ChatGPT-4: Transforming Medical Education and Addressing Clinical Exposure Challenges in the Post-pandemic Era. *Indian Journal of Orthopaedics, 57*(9), 1527–1544.

Maleki, N., Padmanabhan, B., & Dutta, K. (2024). AI Hallucinations: A Misnomer Worth Clarifying. In *Proceedings - 2024 IEEE Conference on Artificial Intelligence, CAI 2024* (pp. 133–138).

Mehandru, N., Miao, B. Y., Almaraz, E. R., Sushil, M., Butte, A. J., & Alaa, A. (2024). Evaluating large language models as agents in the clinic. *npj Digital Medicine, 7*(1), 84.

Meng, X., Yan, X., Zhang, K., Liu, D., Cui, X., Yang, Y., Zhang, M., Cao, C., Wang, J., Wang, X., Gao, J., Wang, Y.-G.-S., Ji, J.-M., Qiu, Z., Li, M., Qian, C., Guo, T., Ma, S., Wang, Z., … Tang, Y.-D. (2024). The application of large language models in medicine: A scoping review. *iScience, 27*(5), 109713.

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.

Miao, J., Thongprayoon, C., Suppadungsuk, S., Krisanapan, P., Radhakrishnan, Y., & Cheungpasitporn, W. (2024). Chain of Thought Utilization in Large Language Models and Application in Nephrology. *Medicina (Lithuania), 60*(1), 148.

Nazi, Z. A., & Peng, W. (2024). Large Language Models in Healthcare and Medical Domain: A Review. *Informatics, 11*(3), 57.

Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. *npj Digital Medicine, 6*(1), 195.

Park, S. H., Suh, C. H., Lee, J. H., Kahn, C. E., & Moy, L. (2024). Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM). *Korean Journal of Radiology, 25*(10), 865–868.

Pfohl, S. R., Cole-Lewis, H., Sayres, R., Neal, D., Asiedu, M., Dieng, A., Tomasev, N., Rashid, Q. M., Azizi, S., Rostamzadeh, N., McCoy, L. G., Celi, L. A., Liu, Y., Schaekermann, M., Walton, A., Parrish, A., Nagpal, C., Singh, P., Dewitt, A., … Singhal, K. (2024). A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine, 30*(12), 3590–3600.

Qiu, J., Lam, K., Li, G., Acharya, A., Wong, T. Y., Darzi, A., Yuan, W., & Topol, E. J. (2024). LLM-based agentic systems in medicine and healthcare. *Nature Machine Intelligence, 6*(12), 1418–1420.

Qiu, J., Li, L., Sun, J., Peng, J., Shi, P., Zhang, R., Dong, Y., Lam, K., Lo, F. P.-W., Xiao, B., Yuan, W., Wang, N., Xu, D., & Lo, B. (2023). Large AI Models in Health Informatics: Applications, Challenges, and the Future. *IEEE Journal of Biomedical and Health Informatics, 27*(12), 6074–6087.

Rauniyar, A., Hagos, D. H., Jha, D., Hakegard, J. E., Bagci, U., Rawat, D. B., & Vlassov, V. (2024). Federated Learning for Medical Applications: A Taxonomy, Current Trends, Challenges, and Future Research Directions. *IEEE Internet of Things Journal, 11*(5), 7374–7398.

Rosenberg, G. S., Magnéli, M., Barle, N., Kontakis, M. G., Müller, A. M., Wittauer, M., Gordon, M., & Brodén, C. (2024). ChatGPT-4 generates orthopedic discharge documents faster than humans maintaining comparable quality: a pilot study of 6 cases. *Acta Orthopaedica, 95*, 152–156.

Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare (Switzerland), 11*(6), 887.

Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y., & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine, 7*(1), 258.

Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., & Gerstein, M. (2024). MEDAGENTS: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (pp. 599–621).

Thirunavukarasu, A. J. (2023). Large language models will not replace healthcare professionals: curbing popular fears and hype. *Journal of the Royal Society of Medicine, 116*(5), 181–182.

Thirunavukarasu, A. J., Mahmood, S., Malem, A., Foster, W. P., Sanghera, R., Hassan, R., Zhou, S., Wong, S. W., Wong, Y. L., Chong, Y. J., Shakeel, A., Chang, Y.-H., Tan, B. K. J., Jain, N., Tan, T. F., Rauz, S., Ting, D. S. W., & Ting, D. S. J. (2023). Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study. *PLOS Digital Health, 3*(4), e0000341.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine, 29*(8), 1930–1940.

Tripathi, S., Sukumaran, R., & Cook, T. S. (2024). Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care. *Journal of the American Medical Informatics Association, 31*(6), 1436–1440.

Wan, P., Huang, Z., Tang, W., Nie, Y., Pei, D., Deng, S., Chen, J., Zhou, Y., Duan, H., Chen, Q., & Long, E. (2024). Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nature Medicine, 30*(10), 2878–2885.

Wang, C., Ong, J., Wang, C., Ong, H., Cheng, R., & Ong, D. (2024). Potential for GPT Technology to Optimize Future Clinical Decision-Making Using Retrieval-Augmented Generation. *Annals of Biomedical Engineering, 52*(5), 1115–1118.

Wang, D., & Zhang, S. (2024). Large language models in medical and healthcare fields: applications, advances, and challenges. *Artificial Intelligence Review, 57*(11), 299.

Williams, C. Y. K., Miao, B. Y., Kornblith, A. E., & Butte, A. J. (2024). Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. *Nature Communications, 15*(1), 8236.

Yigci, D., Eryilmaz, M., Yetisen, A. K., Tasoglu, S., & Ozcan, A. (2025). Large Language Model-Based Chatbots in Higher Education. *Advanced Intelligent Systems, 7*(3), 2400429.

Yildirim, N., Richardson, H., Wetscherek, M. T., Bajwa, J., Jacob, J., Pinnock, M. A., Harris, S., de Castro, D. C., Bannur, S., Hyland, S. L., Ghosh, P., Ranjit, M., Bouzid, K., Schwaighofer, A., Pérez-García, F., Sharma, H., Oktay, O., Lungren, M., Alvarez-Valle, J., … Thieme, A. (2024). Multimodal Healthcare AI: Identifying and Designing Clinically Relevant Vision-Language Applications for Radiology. In *Conference on Human Factors in Computing Systems - Proceedings*. Article 444.

Yu, P., Xu, H., Hu, X., & Deng, C. (2023). Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration. *Healthcare (Switzerland), 11*(20), 2776.

Zhang, C., Xu, J., Tang, R., Yang, J., Wang, W., Yu, X., & Shi, S. (2023). Novel research and future prospects of artificial intelligence in cancer diagnosis and treatment. *Journal of Hematology and Oncology, 16*(1), 114.

Zhang, P., & Kamel Boulos, M. N. (2023). Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges. *Future Internet, 15*(9), 286.

© 2025 by the authors; licensee Growing Science, Canada. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).