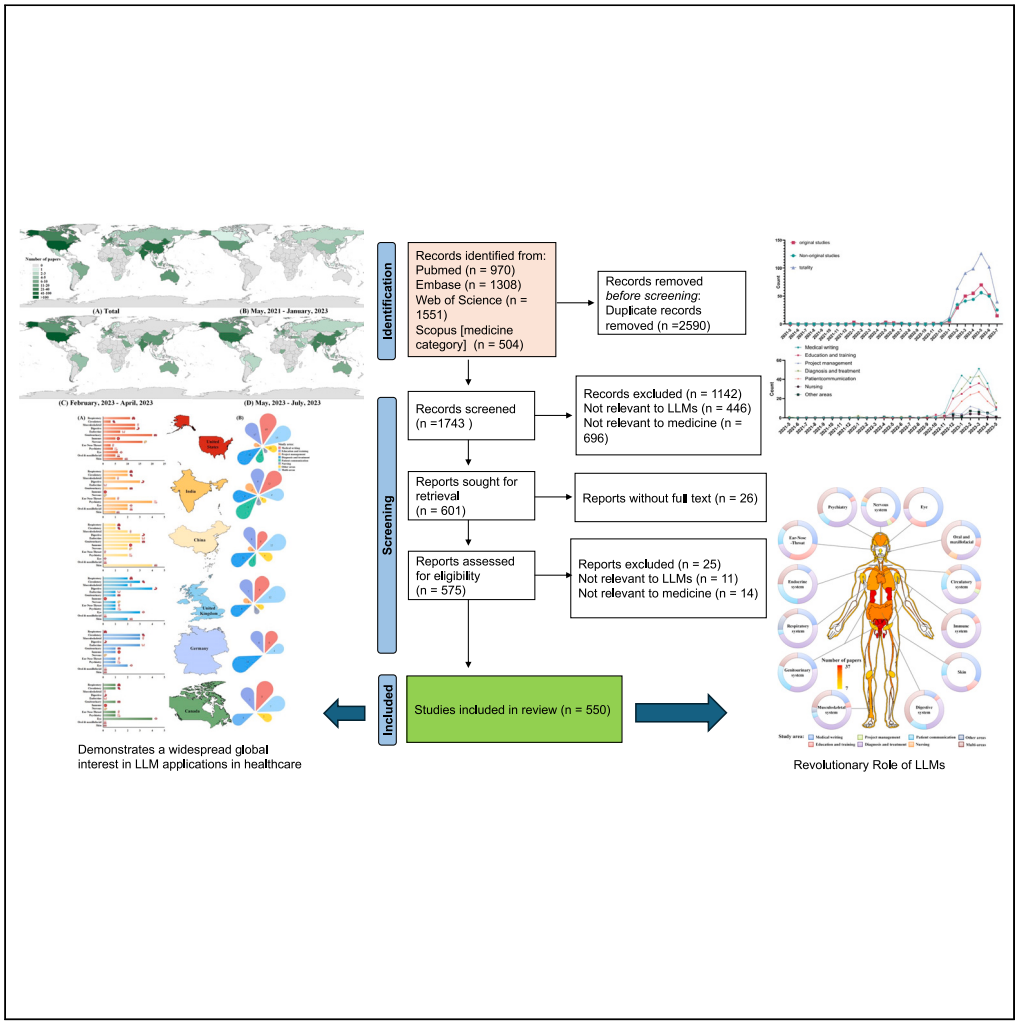
iScience

CellPress
OPEN ACCESS

Article

The application of large language models in medicine: A scoping review



Xiangbin Meng, Xiangyu Yan, Kuo Zhang, ..., Wenyao Wang, Haojun Fan, Yi-Da Tang

fanhj@tju.edu.cn (H.F.)
tangyida@bjmu.edu.cn (Y.-D.T.)

Highlights

LLMs transform healthcare in diagnostics, writing, and education

Multimodal LLMs show great future potential in healthcare

Global surge in LLM research for healthcare applications

Need for ethical LLM integration and empirical studies in clinics

Meng et al., iScience 27, 109713
May 17, 2024
https://doi.org/10.1016/j.isci.2024.109713

iScience

CellPress
OPEN ACCESS

Article

The application of large language models in medicine: A scoping review

Xiangbin Meng,[1,2,13] Xiangyu Yan,[3,13] Kuo Zhang,[4,13] Da Liu,[5] Xiaojuan Cui,[6] Yaodong Yang,[7] Muhan Zhang,[7] Chunxia Cao,[3] Jingjia Wang,[1] Xuliang Wang,[1] Jun Gao,[1] Yuan-Geng-Shuo Wang,[1] Jia-ming Ji,[7] Zifeng Qiu,[8] Muzi Li,[9] Cheng Qian,[1] Tianze Guo,[10] Shuangquan Ma,[11] Zeying Wang,[12] Zexuan Guo,[10] Youlan Lei,[10] Chunli Shao,[1] Wenyao Wang,[1] Haojun Fan,[3,*] and Yi-Da Tang[1,2,14,*]

SUMMARY

This study systematically reviewed the application of large language models (LLMs) in medicine, analyzing 550 selected studies from a vast literature search. LLMs like ChatGPT transformed healthcare by enhancing diagnostics, medical writing, education, and project management. They assisted in drafting medical documents, creating training simulations, and streamlining research processes. Despite their growing utility in assisted diagnosis and improving doctor-patient communication, challenges persisted, including limitations in contextual understanding and the risk of over-reliance. The surge in LLM-related research indicated a focus on medical writing, diagnostics, and patient communication, but highlighted the need for careful integration, considering validation, ethical concerns, and the balance with traditional medical practice. Future research directions suggested a focus on multimodal LLMs, deeper algorithmic understanding, and ensuring responsible, effective use in healthcare.

INTRODUCTION

Large language models (LLMs) have marked a significant milestone in computational linguistics and have rapidly integrated into diverse sectors, including healthcare.[1–3] These models, developed on deep learning and natural language processing technologies, excel in understanding and generating human language.[2–4] By analyzing vast textual datasets, LLMs have gained remarkable capabilities in vocabulary, grammar, semantics, and domain-specific knowledge, especially in medicine.[3–5]

In healthcare, LLMs are revolutionizing diagnostic processes and therapeutic decision-making.[6–11] Their ability to parse through extensive medical records, interpret complex clinical data, and generate meaningful insights is particularly valuable.[12–18] This includes aiding in diagnoses, enhancing medical education, and assisting in research and literature reviews.[19–27] However, while the broad applications of LLMs in medicine are well-documented, there remains a gap in understanding their specific impact on different medical conditions and scenarios.[28–34]

This research aims to fill this gap by providing a comprehensive systematic review of LLM applications in medicine. We focus on delineating the landscape, identifying challenges, and exploring future possibilities. By condensing the vast knowledge into actionable insights, we hope to guide further advancements in this intersection of artificial intelligence (AI) and healthcare.

[1]Department of Cardiology and Institute of Vascular Medicine, Peking University Third Hospital, Beijing, China
[2]State Key Laboratory of Vascular Homeostasis and Remodeling, Peking University, Beijing, China
[3]Institute of Disaster and Emergency Medicine, Tianjin University, Tianjin, China
[4]Department of Cardiology, State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences, and Peking Union Medical College, Beijing, China
[5]Department of Cardiology, the First Hospital of Hebei Medical University, Graduate School of Hebei Medical University, Shi-jia-zhuang, Hebei, China
[6]School of Software & Microelectronics, Peking University, Beijing, China
[7]Institute for Artificial Intelligence, Peking University, Beijing, China
[8]Peking University Health Science Center, Peking University First Hospital, Beijing, China
[9]Peking University Health Science Center, Peking University People's Hospital, Beijing, China
[10]Peking University Health Science Center, Beijing, China
[11]School of Biological Science and Medical Engineering, Beijing Advanced Innovation Centre for Biomedical Engineering, Beihang University, Key Laboratory for Biomechanics and Mechanobiology of Ministry of Education, Beijing, China
[12]Department of Prosthodontics, Peking University School and Hospital of Stomatology, National Engineering Laboratory for Digital and Material Technology of Stomatology, National Clinical Research Center for Oral Diseases, Beijing Key Laboratory of Digital Stomatology, Beijing, China
[13]These authors contributed equally
[14]Lead contact
*Correspondence: fanhj@tju.edu.cn (H.F.), tangyida@bjmu.edu.cn (Y.-D.T.)
https://doi.org/10.1016/j.isci.2024.109713

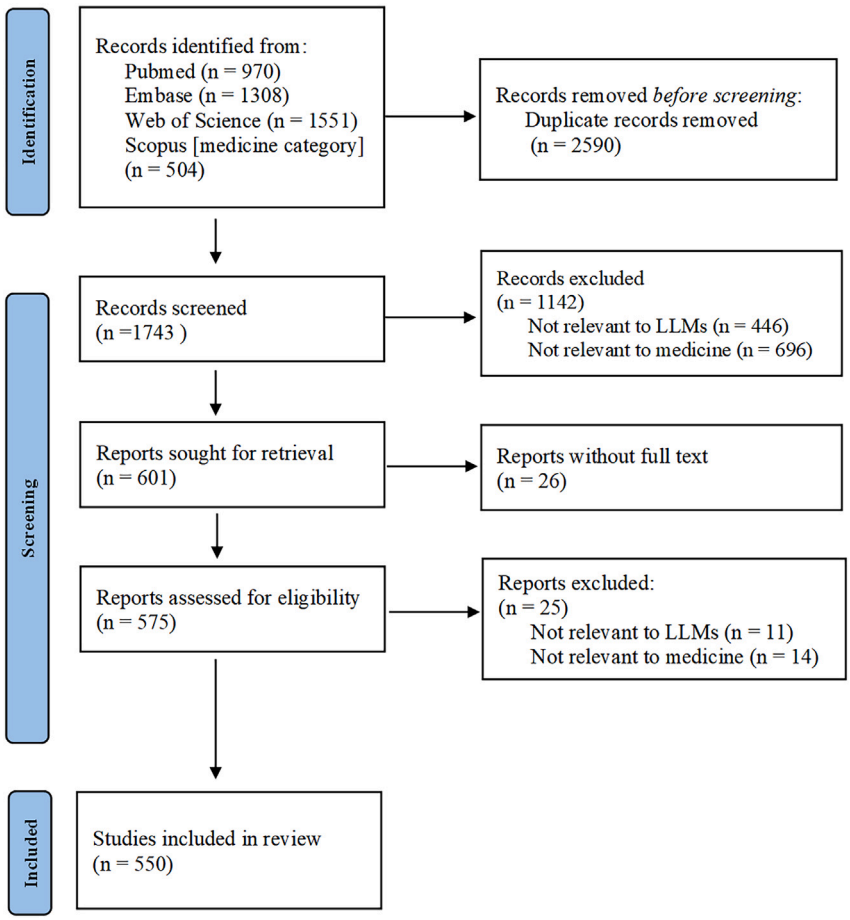


Figure 1. Search results and study inclusion

Figure360▷ For a Figure360 author presentation of Figure 1, see https://doi.org/10.1016/j.isci.2024.109713.

RESULTS

Search results and study inclusion

Our meticulous search across renowned databases such as PubMed, EmBase, Web of Science, and Scopus (specifically the "medicine" category) yielded a total of 4,323 articles up until 20 July 2023. After eliminating 2,590 duplicate articles, 1,743 records were earmarked for potential inclusion. A rigorous selection process subsequently ruled out 1,193 articles for various reasons, including irrelevance to LLMs, misalignment with the medical domain, and a lack of accessible full text. Consequently, a final count of 550 studies was selected for the scoping review (Table S1) (Figure 1).

The rapid increase of LLM-related medical researches

Our data presented a compelling trajectory of LLM-related medical research publications from May 2021 to July 2023. Initially, in May 2021, there was a single publication, Korngiebel and Mooney, delving into the burgeoning realm of natural language computer applications, specifically generative pretrained transformer 3 (GPT-3). Highlighting its potential to revolutionize traditional human-dominated healthcare interactions, we make a cautionary note. They emphasized the imperative to judiciously weigh the advantages and challenges before such sophisticated tools find footing in the intricate tapestry of healthcare delivery.[35] As the horizon of eHealth broadens, the paper serves as a timely reminder that as innovation becomes more prevalent, its integration into the clinical milieu requires circumspection and preparation.

A discernible increase was evident in November and December 2022, with three and four publications, respectively. The year 2023 shows a significant surge, particularly from January to June. May 2023 has the highest number of publications (126 articles). Notably, there was a sharp decline in July 2023 to only 40 articles. This dip can be attributed to the inherent lag in the database update process. Typically, accepted articles take approximately two months from initial acceptance to database inclusion (Figure 2).

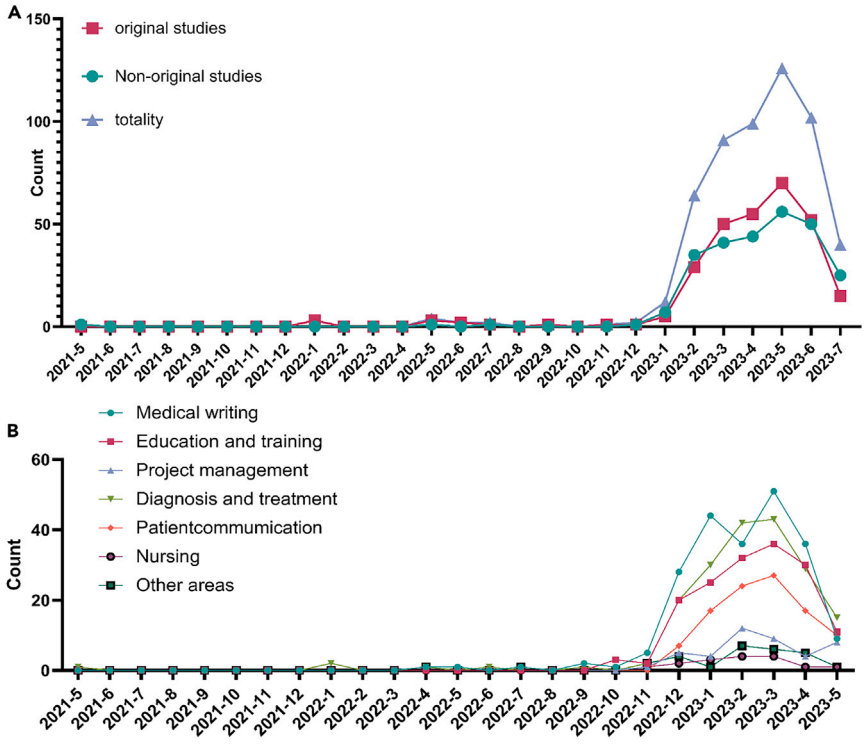


Figure 2. Monthly breakdown of LLMs medical research: original vs. non-original publications

Figure360▷ For a Figure360 author presentation of Figure 2, see https://doi.org/10.1016/j.isci.2024.109713.

The role of LLMs in transforming medical practices

The integration of LLMs in medical writing

With the emergence of LLMs, there has been a clear revolution in the field of medical writing, particularly the incorporation of LLMs such as ChatGPT.[36] These models' evolving capabilities have made their integration into medical documentation increasingly relevant. While LLMs are not without limitations in comprehending and generating medical texts, their proficiency in swiftly accessing a breadth of multidisciplinary data is noteworthy. They efficiently retrieve up-to-date research from medical literature, aiding researchers in rapidly synthesizing new findings.[37] Notably, LLMs excel in drafting initial versions of articles and refining grammar and style in existing documents, thus enhancing their clarity and coherences.[38,39] In particular, ophthalmologists have utilized ChatGPT for swiftly creating summaries and notes for ophthalmic surgeries.[40,41] Additionally, LLMs can simplify complex medical statistics for better comprehension.[42] LLMs also contribute to the linguistic improvement of document manuscripts.[43] When researchers need to explain complex medical data or statistical results, LLMs can provide clear and concise explanations, making them easier for readers to understand.[44] However, their limitations, such as occasionally citing non-existent references, must be acknowledged. Despite these challenges, the potential for LLMs in medical writing remains substantial.

To fully leverage LLMs in medical writing, enhancing their contextual grasp of medical subtleties, incorporating fact-checking features, adapting to various medical writing styles, and establishing a feedback loop with healthcare professionals is essential.[45,46] Crucial steps include ensuring transparency in authorship, integrating with credible medical databases, prioritizing data security, and encouraging collaboration between developers and medical experts for effective and trustworthy LLM applications in healthcare (Figure S1).

LLMs in medical examinations and education

LLMs, such as ChatGPT, have shown considerable potential in enhancing medical education and examination practices. These models offer innovative tools for medical training, including realistic patient-doctor role-playing simulations with real-time feedback. They enable the creation of diverse medical scenarios, complete with detailed patient histories and decision-making processes.[47,48] This advancement aids in improving the training efficiency of medical personnel, thereby enhancing the skills and capabilities of future healthcare professionals. Additionally, LLMs can generate new practice questions from existing exam databases, offering students a broader range of study materials.[49] Their ability to automatically score objective questions, such as multiple-choice or true-or-false formats, and provide preliminary evaluations for short-answer and essay responses, underscores their utility in educational settings.[50,51] However, it's crucial to recognize that while LLMs offer highly accurate responses, they are not flawless. Supervision is necessary to prevent misinformation and ensure that medical students receive accurate and reliable guidance.

Despite the advancements brought by LLMs in medical training and exam preparation, concerns about their limitations, including the potential spread of inaccuracies and the risk of over-reliance by students, must be addressed.[52] To effectively utilize LLMs in medical education, integrating expert-reviewed content is vital. Additionally, embedding mechanisms for continuous learning and rectification of inaccuracies, coupled with emphasizing human oversight, is essential. These steps will ensure that LLMs act as supplemental tools, reinforcing rather than replacing traditional learning methods[53] (Figure S1).

The synergy of LLMs in medical project and research management

In the fields of medical project management and research design, LLMs have emerged as pivotal tools, enhancing a range of processes including literature searches, project planning, risk assessment, ethics review, data management, budgeting, and team collaboration.[37] For instance, LLMs can swiftly distill key insights from extensive literature, laying a robust foundation for research initiatives.[54] They also have the capability to anticipate potential project risks, alerting teams to these issues early and proposing mitigation strategies. In data management, LLMs adeptly organize and categorize vast quantities of medical data, streamlining this often complex process.[46] Regarding budget and funding management, they generate budget proposals based on project needs and historical trends, aiding in the judicious allocation of resources. Crucially, LLMs enhance communication and collaboration within research teams, fostering stronger connections between researchers, patients, and the wider audience, thereby ensuring that research aligns with real-world needs.[55] The expectation is that LLMs will alleviate the burden of administrative tasks and paperwork in research planning, offering significant time savings for medical staff. Nonetheless, the accuracy of LLM outputs warrants close scrutiny, and challenges arise in addressing language and cultural variances in their application (Figure S1).

In the context of medical project management and research design, the implementation of LLMs promises substantial efficiency gains, from optimizing literature reviews to improving collaborative efforts. However, their integration is not without challenges, notably in maintaining data accuracy and addressing linguistic and cultural subtleties. As LLMs begin to transform these administrative domains, the importance of rigorous validation protocols, cultural contextualization, and continual updates with the latest medical data becomes paramount.[56] Such measures will ensure that LLMs are both efficient and reliable, serving as invaluable aids in medical research and project management without compromising their integrity.

Revolutionizing medical auxiliary diagnosis with LLMs

LLMs, exemplified by tools like ChatGPT, are revolutionizing the landscape of assisted diagnosis in healthcare. These AI-driven systems adeptly integrate clinical nuances with advanced algorithms, showcasing their proficiency in various medical fields, from the complexities of gastroenterology to the precision of dentistry.[57–59] Beyond auxiliary diagnostics, LLMs extend their utility to recommending medical examinations, providing literature support, and enhancing doctor-patient interactions.[60] Given the stringent standards in healthcare for accuracy, safety, reliability, and clarity, the application of such tools in medical decision-making is subject to rigorous scrutiny. Medical decision-making is inherently intricate, often involving the interpretation of ambiguous, unstructured data, and multifaceted medical knowledge. LLMs' limitations are evident in their struggles with contextual understanding and causal reasoning.[61,62] Instances where LLMs offer misleading or potentially hazardous advice underscore the need for cautious adoption in clinical settings; nevertheless, they hold considerable promise for transformative changes in healthcare. As the medical field navigates the surging currents of digital health, the potential of AI is unmistakable.[63] However, this advancement warrants a prudent approach, with a consensus in the medical community that AI should augment, not supplant, human expertise.

In summary, while LLMs undoubtedly introduce groundbreaking shifts and novel opportunities in healthcare, their rise is checked by significant challenges. These include the need for more robust evidence-based validation in clinical contexts, an often-superficial grasp of disease pathophysiology favoring probabilistic over deterministic reasoning, and emerging ethical issues, particularly concerning data privacy and transparency. Navigating this AI-integrated medical era necessitates a harmonious blend of technological innovation and steadfast adherence to core medical principles. Emphasizing a balanced, evidence-driven, and collaborative strategy is essential to uphold the integrity and quality of patient care (Figure S1).

Elevating medical communication in the LLMs

In today's complex medical landscape, AI-driven platforms like ChatGPT are significantly altering the dynamics of doctor-patient communication and aiding in the dissemination of medical knowledge. LLMs serve as preliminary online consultation tools, offering patients basic but essential information about their conditions, treatments, and preventive measures.[64,65] This approach can not only save patients time but also equip them with foundational knowledge and advice prior to in-person medical consultations. One of the key strengths of LLMs lies in their ability to demystify medical terminology, providing patients with clear and understandable explanations, thereby enhancing their comprehension of medical diagnoses and recommendations.[66] Furthermore, recent studies underscore the role of these advanced tools in improving consultation accuracy and countering medical misinformation, particularly in areas such as vaccination.[67] The expansive knowledge base of LLMs enables coverage across diverse medical fields, from orthodontics to cardiac surgery, potentially narrowing the communication divide between doctors and patients.[32] However, while these tools mark a significant shift in enhancing doctor-patient interactions, it's crucial to acknowledge their inherent limitations. Therefore, continuous education and training of medical personnel are essential to ensure the responsible and effective use of LLMs in clinical settings.

As LLMs pave the way for breakthroughs in medical communication, bridging gaps and simplifying complex medical concepts, their limitations must be carefully considered. The guidance provided by LLMs, though extensive, may lack the intricate understanding inherent in human medical practice, and there's a risk of misinformation if these tools are not meticulously supervised.[68] Additionally, the inherently

impersonal nature of LLMs can overlook the emotional and psychological aspects crucial in patient care.[69] To fully leverage the benefits of LLMs in doctor-patient communication, it is imperative to refine their algorithms for more profound contextual comprehension, continually update them with clinically validated information, and incorporate a level of human oversight to ensure the accuracy and personalization of their responses (Figure S1).

Innovative LLMs applications in nursing

The advent of AI, exemplified by models like ChatGPT, signals a significant transformation in healthcare, especially in nursing education. These advanced models serve not only as repositories of vast information but also as dynamic tools reshaping how nursing education is delivered. They facilitate more realistic, relevant assessments and prepare nursing students and professionals for the diverse challenges of clinical practice. LLMs, utilizing patient medical records and current conditions, offer personalized nursing recommendations and help develop tailored care plans.[70] They assist nurses in managing and accurately dosing medication, and provide timely reminders to both medical staff and patients. An automated nursing record feature streamlines workflow, enhancing the accuracy and completeness of patient information. For those in rural or remote areas, LLMs offer essential nursing advice and support, ensuring timely assistance is available.[71] Additionally, these models can offer emotional support, aiding patients in managing the psychological stresses associated with illness.[72] Despite these benefits, it is crucial to recognize that such technologies cannot fully replace the irreplaceable value of human care and interaction.[73] Concerns such as safeguarding patient privacy, preventing overdependence on technology, and preserving the authenticity of human interactions remain paramount. As we integrate these technologies into nursing practice, it is imperative to use them judiciously, ensuring they complement rather than replace the fundamental elements of human interaction and care (Figure S1).

The applications of LLMs in other research areas

In the multifaceted realm of biomedicine, LLMs are showcasing the extensive capabilities of AI. A prime example is in literature retrieval and analysis, where LLMs efficiently collate and synthesize information from a vast array of biomedical publications.[74] This capability is invaluable to researchers, enabling rapid access to and organization of the latest findings on specific drugs, diseases, or genetic research.[75] In the arena of drug discovery, LLMs aid in predicting the activity, toxicity, and pharmacokinetic properties of new drug compounds, offering significant time savings and early-stage screening potential.[76] Additionally, in genomics, LLMs contribute to annotating the functions of newly identified genes using existing literature and databases, a crucial step given the daily discovery of new genes.[77] While protein structure prediction primarily relies on specialized models like AlphaFold, LLMs enhance these models by providing supplementary literature-based insights, thus refining the accuracy of such predictions.[78,79] In epidemiology, LLMs are instrumental in tracking and forecasting disease spread by analyzing online textual data, offering substantial support for public health decision-making.[80,81] Furthermore, LLMs find application in bioinformatics for predicting patterns, functional domains, and sequence similarities (Figure S1).

However, despite the wide-ranging utility of LLMs in biomedicine, they cannot entirely supplant laboratory experimentation or deep-rooted biomedical expertise. Particularly, drug molecule predictions made by LLMs may not always align with actual biological responses. Gene function predictions, informed though they may be, might not fully account for complex gene interactions. Similarly, relying solely on LLMs for epidemiological trends could lead to inaccuracies if not cross-validated with actual data. Therefore, while LLMs represent a significant advancement in biomedicine, their optimal use involves a balanced approach. This approach should integrate their computational power with thorough experimental validation and expert analysis to ensure scientific accuracy and maintain the integrity of biomedical research.

Distribution and type of LLM-related medical research across organ systems

A total of 219 (39.82%) articles focused on at least one specific disease across different organ systems. The genitourinary system led the chart with 37 studies, of which 13 focused on assistive diagnostics and 9 on patient-nurse communication. The circulatory and digestive systems both had 27 studies, with the former showing a greater contribution to medical writing (six studies) and the latter leaning toward multi-faceted applications (10 studies). The nervous and psychiatric domains exhibited broad interest in medical writing, with nine and four studies, respectively. In contrast, the ear-nose-throat system had the least representation, with seven studies. Notably, the majority of studies across all systems have concentrated on assistive diagnostics and medical writing (Figure 3).

Geographical distribution of LLM-related medical research publications

The data illustrated the comprehensive geographical distribution of LLM-related medical research across 57 countries. In the initial phase, the publications originated in 12 countries. This expanded to 44 and 47 countries in the second and third phases, respectively. The US led the list with 221 publications, followed by India with 53 publications and China with 49 publications. Other notable contributors included the UK with 37 publications, Germany with 27 publications and Canada with 26 publications. This distribution of research outputs clearly indicates a global interest and contribution in the field of LLM in medicine, encompassing both developed and developing nations, and underscores the widespread relevance of this technology in the medical sector (Figure 4).

An international landscape: Delineating LLM-related medical research by domain and nation

The data provide a meticulous portrayal of scholarly pursuits across distinct medical paradigms, delineated by country. The US has emerged as a clear leader in academic excellence, with significant contributions across various domains: 60 entries in medical writing, 37 in education

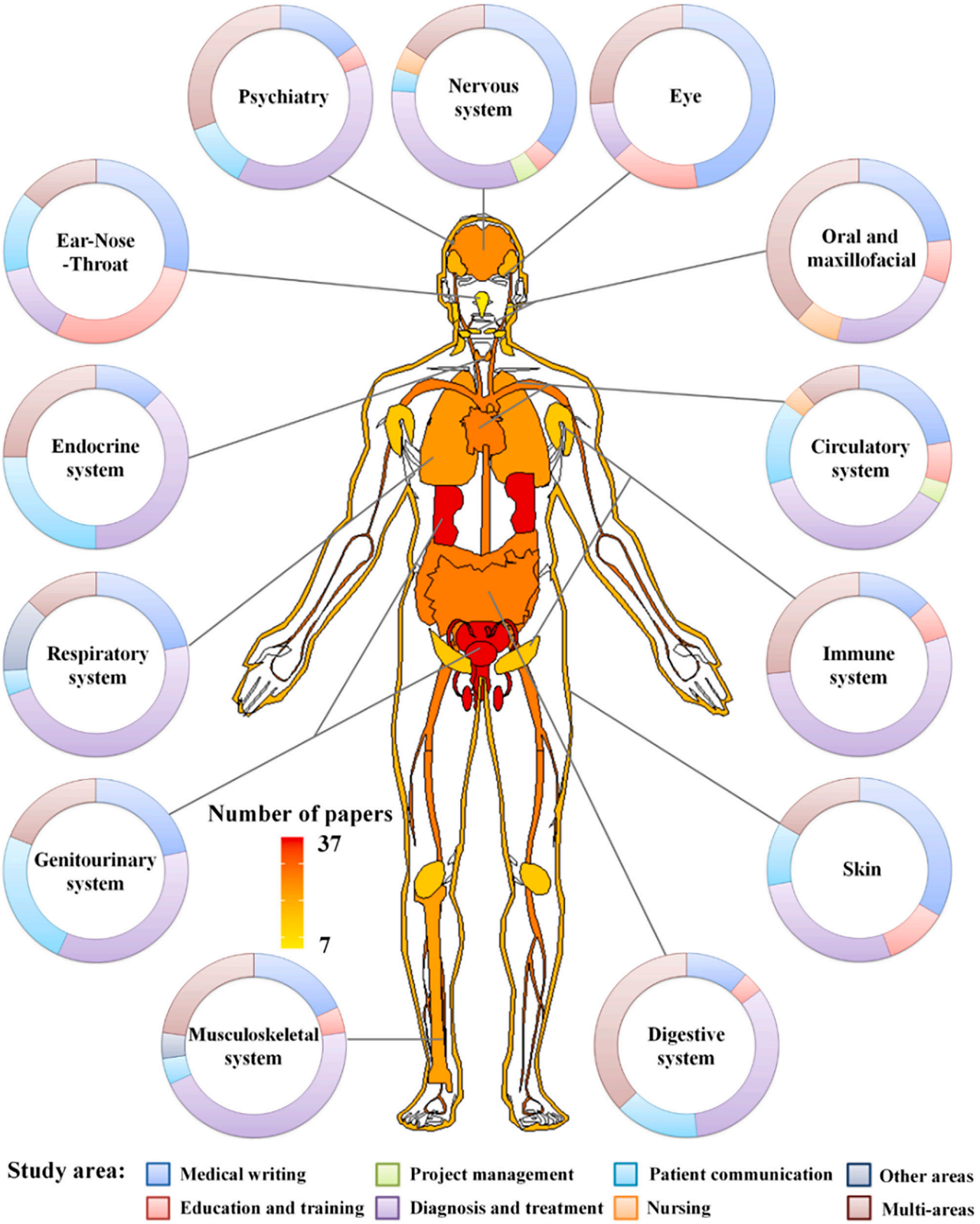


Figure 3. Heatmap of LLMs applications in medical research by organ system

Figure360▷ For a Figure360 author presentation of Figure 3, see https://doi.org/10.1016/j.isci.2024.109713.

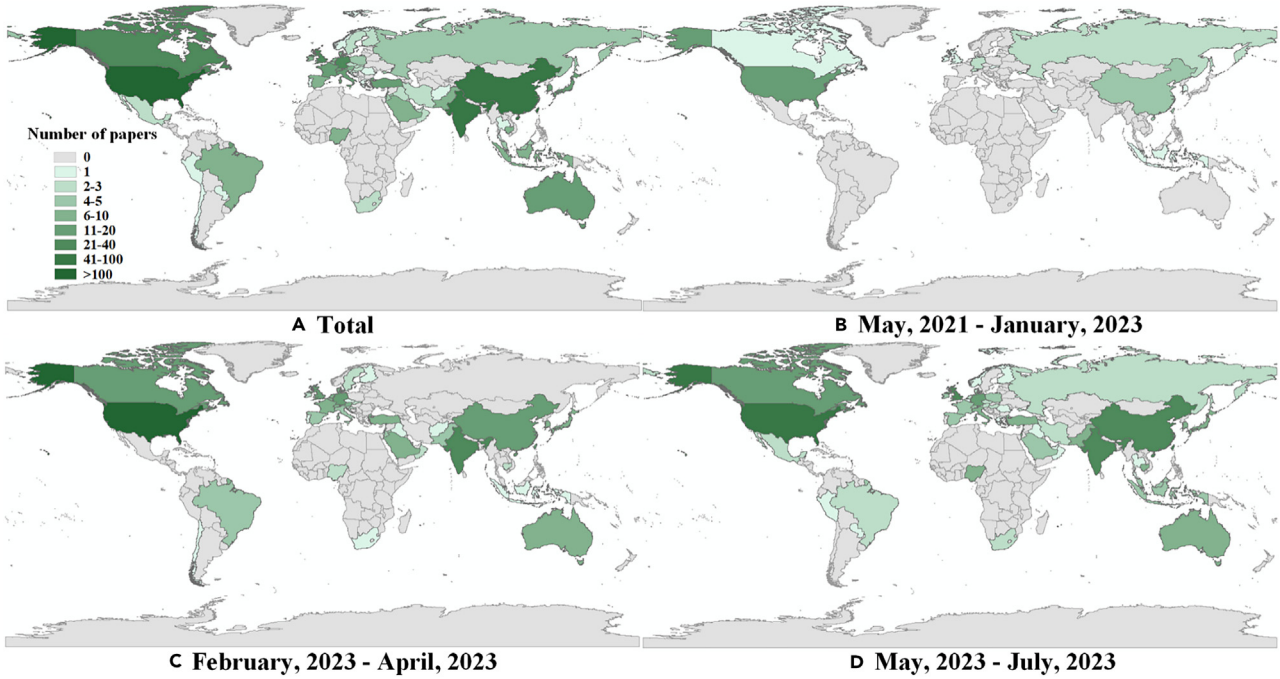


Figure 4. Global spread of LLMs medical research by country

Figure360▷ For a Figure360 author presentation of Figure 4, see https://doi.org/10.1016/j.isci.2024.109713.

and training, 42 in diagnosis and treatment, and 54 in multidisciplinary pursuits. India, while not matching the volume of the US, shows a focused emphasis in medical writing and patient communication, with 12 and 11 entries, respectively. It also maintains a balanced approach across other areas and shows a notable presence in interdisciplinary studies with 8 entries. China, with a conservative count in medical writing, displays considerable progress in interdisciplinary research and diagnosis and treatment. Following these, Germany's academic strength is prominently inclined toward diagnosis and treatment, paralleling the US in its scholarly focus. The UK, with a rich academic heritage, contributes across various fields: 9 entries in medical writing, 7 in pedagogy, and 12 in multidisciplinary research. Lastly, Canada, though moderate in overall numbers, demonstrates a distinct emphasis in multidisciplinary research and ophthalmology.

When viewed through an anatomical lens, the US leads, particularly in musculoskeletal and digestive systems research, and shows notable engagement in genitourinary, respiratory, and neural studies. India aligns its research focus on respiratory and psychiatric areas. China offers balanced contributions, especially in digestive and endocrine systems, while also attending to musculoskeletal and neural research. The UK, with its balanced academic approach, emphasizes circulatory and digestive system studies. Germany, with a diverse academic portfolio, leans toward circulatory and musculoskeletal research. Canada, though modest in its contributions, distinctly focuses on ophthalmological research (Figure 5).

DISCUSSION

This is a pioneering study to quantitatively and comprehensively chart the integration of LLMs into the medical domain. The recent upsurge in the adoption of LLMs can be attributed to the open accessibility of groundbreaking platforms, such as GPT.[1,8,82] A notable observation is the surge in original research, which is indicative of growing empirical endeavors to harness the potential of LLMs in real-world medical scenarios.[83–85]

The geospatial distribution of research shows a pronounced skew toward regions, such as the US, India, and Germany. These regions, with higher internet access, have undeniably provided fertile ground for the intersection of AI and medicine.[86–88] Specifically, the synergy between the burgeoning IT industries and advanced healthcare infrastructure in these countries may explain their dominance in the medical LLM landscape. While the application of LLMs spans various medical disciplines, our findings underscore the inclination toward certain organ systems, notably the genitourinary, digestive, and circulatory systems.[89–92] Research in these areas largely focuses on facets such as medical writing, auxiliary diagnosis, and patient communication, hinting at immediate areas where LLMs, such as ChatGPT, can bring tangible benefits. Concurrently, fields such as psychiatry and the nervous system, although well represented, emphasize uncharted areas where further research could unearth novel applications. The expansive datasets also show an emergent trend: while LLMs have found robust applications in patient communication, medical writing, and auxiliary diagnosis, there remains latent potential in realms such as medical education and training, especially in simulating patient-doctor interactions.[10,85,93–95] In addition, the challenges posed by linguistic and cultural nuances in LLMs underscore the importance of region-specific model training and data integration.

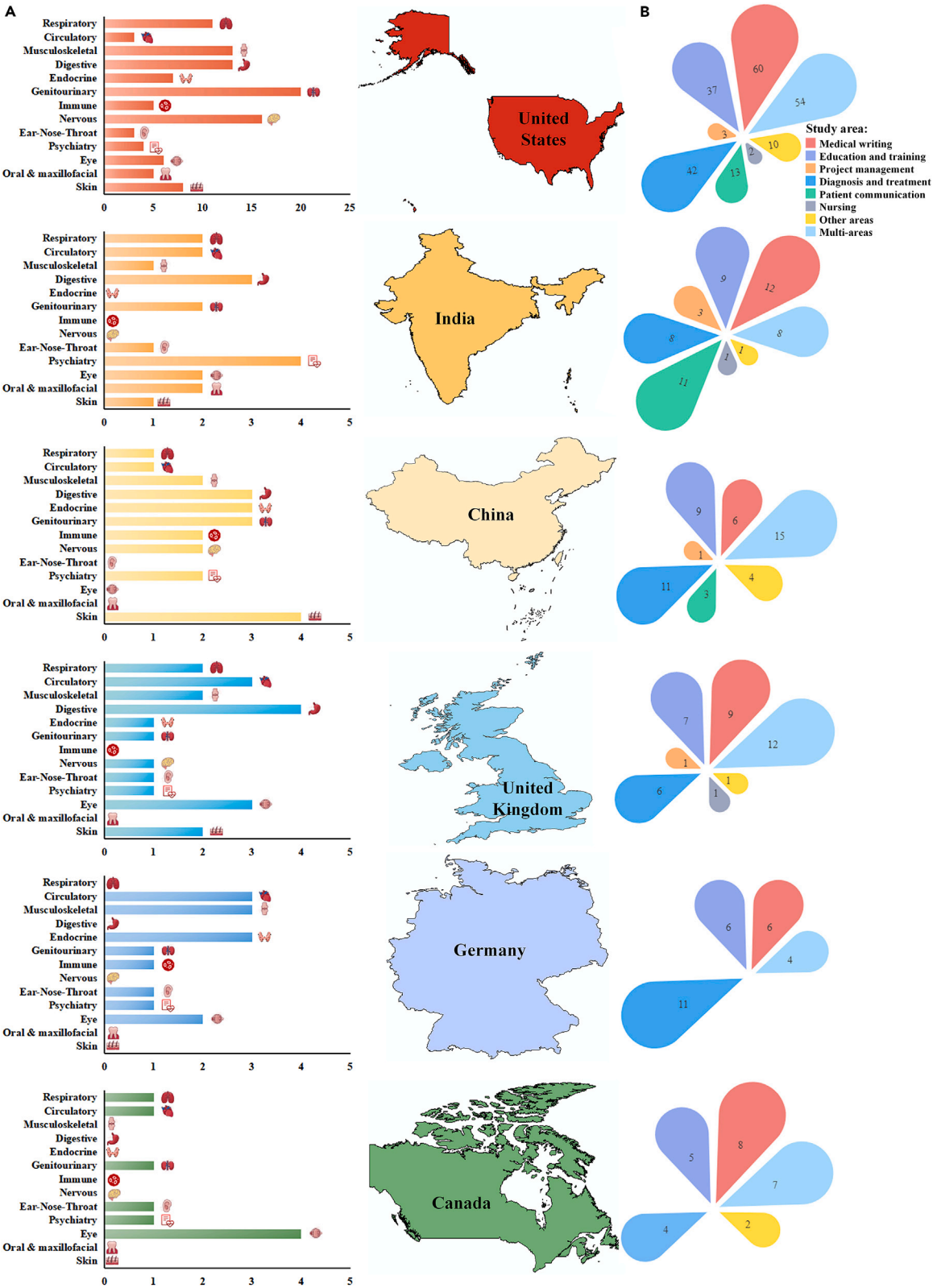

A
Respiratory
Circulatory
Musculoskeletal
Digestive
Endocrine
Genitourinary
Immune
Nervous
Ear-Nose-Throat
Psychiatry
Eye
Oral & maxillofacial
Skin
0 5 10 15 20 25
United States
B
60
37
54
3
10
2
42
13
Study area:
Medical writing
Education and training
Project management
Diagnosis and treatment
Patient communication
Nursing
Other areas
Multi-areas
0 1 2 3 4 5
India
9
12
3
8
8
1
1
11
China
9
6
15
1
11
4
3
United Kingdom
7
9
12
1
1
6
1
Germany
6
6
4
11
Canada
5
8
7
4
2

Figure 5. Comparative visualization of LLMs medical contributions across nations

Figure360▷ For a Figure360 author presentation of Figure 5, see https://doi.org/10.1016/j.isci.2024.109713.

(A) Distribution of disease organ publications in research by country.

(B) Comparative overview of medical field publications by country.

The advent of multimodal LLMs is bringing about a paradigm shift in the medical field, offering the capability to process and generate diverse data types, thereby unlocking unprecedented potential.[96–99] To understand their role, it's essential to define what multimodal LLMs entail. These models are adept at processing, interpreting, and generating a wide array of data types, including but not limited to text, images, sounds, and more. This versatility implies that within the medical arena, LLMs are equipped to handle not only textual data, such as patient medical records, diagnostic reports, and research papers, but also to interpret and analyze medical imaging data like MRIs, CT scans, and X-rays, as well as other forms of data like voice recordings or biomarkers. Multimodal LLMs offer a novel and efficacious approach to diagnosis, treatment, and healthcare management. Their robust capabilities in data processing and integration allow medical professionals to deliver more precise and efficient services to patients. At the same time, these models enable patients to access medical advice and care with greater convenience. As these technologies continue to evolve and improve, their significance and impact in the medical field are expected to grow exponentially.

Traditional machine learning systems (MLS), including random forests and traditional neural networks, use specific structured data to improve the identification of critically ill patients.[100] MLS employs a predictive model and provides real-time explanations via the shapley additive explanations (SHAP) method to help medical staff understand why certain patients may require immediate treatment. In contrast, LLMs like GPT-4 are primarily designed to understand and generate natural language. LLMs offer flexibility in handling a wide variety of data, including unstructured data, but they do not inherently provide the same structured, real-time interpretation as MLS. Traditional MLS faces challenges such as overfitting, lack of interpretability, and difficulty in handling diverse and unstructured data. LLMs, with their flexibility, have the potential to bridge these gaps, making them more valuable in clinical applications, especially when patient data are diverse and unstructured.

The predictive capabilities of LLMs stem from their pre-training on vast datasets, enabling them to understand and generate language, and perform a variety of predictive tasks including text classification, question answering, and summary generation.[101] Unlike MLS, LLMs, through their deep neural network architecture, can automatically identify and leverage complex patterns and relationships within language data, without the need for pre-selecting variables or explicit feature engineering.[41,102] LLMs are capable of processing and analyzing large volumes of unstructured text data, utilizing knowledge learned from pre-trained models to execute complex reasoning tasks and predictions.[103,104] This means that LLMs can recognize and utilize complex patterns and associations that MLS might overlook, thereby providing more accurate predictions in certain cases.[105] In fact, LLMs can achieve all the predictive capabilities constructed through deep learning or logistic regression modeling by MLS, and possess a strong generalization ability. MLS struggles with missing or anomalous input variables, leading to ineffective operation or stability issues, whereas LLMs are not limited by fixed input variables, demonstrating superior intelligent processing capabilities and closer alignment with human thought processes.[106] Of course, both MLS and LLMs face challenges related to data bias and model transparency. Future research directions need to continuously improve model design, optimization algorithms, and training methods to address these issues.[107,108]

The utilization of LLMs in medicine, as presented, undeniably brings transformative prospects to healthcare, from diagnostics to patient communication. However, the juxtaposition of promise with persistent challenges warrants a thoughtful discourse. A pressing concern lies in the validation of LLMs in real-world clinical settings. While LLMs have showcased potential, their clinical efficacy remains largely untested. Ensuring these models align with evidence-based medicine standards becomes crucial, requiring rigorous studies that measure their reliability against established clinical benchmarks.

The integration of LLMs into the medical field introduces a significant challenge: the current scarcity of evidence-based medical research concerning the application of LLMs in healthcare settings.[109] Although LLMs have shown remarkable efficacy in various sectors, the unique context of medicine, with its direct implications for human life and health, necessitates a cautious approach to the introduction of untested technologies or methods into clinical practice.[110] Despite their robust data processing capabilities, LLMs present a potential risk for prediction errors in clinical environments. The medical domain, with its complex interplay of biology, physiology, and pathology, might be challenging for machine learning models to fully encapsulate, especially considering the intricacies and variability inherent in medical data.[111] Furthermore, the realm of medical decision-making often requires a high level of expertise and experience, aspects that may not be entirely replicable by LLMs. The consequences of medical decisions far surpass those in other sectors, where a misdiagnosis or incorrect treatment recommendation could directly jeopardize a patient's life. Hence, it is imperative to back any new technological innovation, including LLMs, with solid scientific evidence before they are implemented in medical practice.

Currently, empirical studies examining the application of LLMs in the medical field are limited. This scarcity of research implies an inability to definitively assess the accuracy, reliability, and safety of LLMs within a healthcare context. To comprehensively understand the potential benefits and risks associated with LLMs in medicine, a more robust body of clinical research is required. This research should encompass randomized controlled trials, observational studies, and extensive collaborative research, which are critical to evaluate the clinical utility of LLMs accurately.[32] Challenges such as data privacy, model interpretability, and potential biases persist. Ensuring data security, providing clear model outputs, and addressing biases are crucial for LLMs' success in medicine.[65] The development of these models must balance technical performance with societal trust and ethical considerations in healthcare. The American Medical Association's recent principles for AI in medicine highlight the need for a comprehensive evaluation system for LLMs in medical scenarios.[112] This system should assess not only accuracy

and efficiency but also safety, applicability, and impact on patients. It requires collaboration among medical experts, computer scientists, regulatory bodies, and the public to ensure LLMs are used responsibly and effectively. To fully realize the benefits of LLMs in healthcare, efforts must be made to make these technologies accessible and inclusive, overcoming language, cultural, and technological barriers. Training for medical professionals and public education are essential to improve medical services and promote health and well-being.

In conclusion, the integration of LLMs such as GPT-4 into the medical field represents a significant advancement with the potential to enhance healthcare in numerous ways, including diagnostics, patient communication, and medical training. These models' ability to process diverse data types makes them valuable in the evolving digital healthcare landscape. However, this potential is balanced by the need for rigorous clinical validation and testing to ensure they meet evidence-based medicine standards without compromising patient care. Ethical considerations, such as data privacy, potential biases, and the risk of over-reliance on AI, are crucial, emphasizing that LLMs should augment rather than replace human medical expertise. Addressing these challenges requires a collaborative approach involving medical professionals, AI researchers, regulatory bodies, and patients, aiming to develop LLMs that are technically effective, ethically sound, and socially responsible. Recognizing LLMs' limitations in understanding complex medical conditions and human nuances is vital, and future development should focus on bridging these gaps to supplement the irreplaceable human elements of empathy, ethical judgment, and clinical intuition in healthcare. This balanced approach is essential as we embrace AI's role in medicine, focusing on improving patient outcomes and healthcare delivery while responsibly navigating the ethical, practical, and clinical challenges.

Limitations of the study

This study, while comprehensive, has several limitations. Firstly, our reliance on selected databases might have excluded pertinent publications from non-indexed sources. The dynamic nature of AI and LLM research also means that newer advancements post-July 2023 are not captured. Since relevant clinical research is still in its infancy, this article currently lacks an application summary of specific scenarios.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Study design
  - Literature search
  - Study selection
  - Data extraction
  - Data integration

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109713.

ACKNOWLEDGMENTS

This study was funded by the National Key R&D Program of China (2020YFC2004705), National Natural Science Foundation of China (81825003, 91957123, and 82270376), CAMS Innovation Fund for Medical Sciences (2022-I2M-C&T-B-119 and 2021-I2M-5-003), Beijing Nova Program (Z201100006820002) from Beijing Municipal Science & Technology Commission, and CSC Special Fund for Clinical Research (CSCF2021A04).

AUTHOR CONTRIBUTIONS

X.M., X.Y., and K.Z. (equal contribution) were responsible for the study's concept and design, analysis and interpretation of data, statistical analysis, and drafting of the manuscript. D.L., X.C., Y.Y., M.Z., C.C., J.W., X.W., J.G., Y.-g.-s.W., J.-m.J., Z.Q., M.L., C.Q., T.G., S.M., Z.W., Z.G., Y.L., C.S., and W.W. contributed to the research performance, acquisition, and analysis of data, and participated in drafting the manuscript. H.F. and Y.-D.T. (shared senior authorship) provided supervision, project administration, and funding acquisition. They also contributed to the study concept and design, provided administrative support, and critically revised the manuscript for important intellectual content. All authors have read and approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 16, 2023
Revised: January 27, 2024
Accepted: April 7, 2024
Published: April 23, 2024

REFERENCES

1. Minssen, T., Vayena, E., and Cohen, I.G. (2023). The Challenges for Regulating Medical Use of ChatGPT and Other Large Language Models. JAMA *330*, 315–316. https://doi.org/10.1001/jama.2023.9651.
2. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A Survey of Large Language Models. Preprint at arxiv. https://doi.org/10.48550/arXiv.2303.18223.
3. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. Preprint at arxiv. https://doi.org/10.48550/arXiv.2203.02155.
4. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural Inf. Process. Syst. *35*, 24824–24837.
5. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent Abilities of Large Language Models. Preprint at arxiv. https://doi.org/10.48550/arXiv.2206.07682.
6. Azizi, Z., Alipour, P., Gomez, S., Broadwin, C., Islam, S., Sarraju, A., Rogers, A., Sandhu, A.T., and Rodriguez, F. (2023). Evaluating Recommendations About Atrial Fibrillation for Patients and Clinicians Obtained From Chat-Based Artificial Intelligence Algorithms. Circ. Arrhythm. Electrophysiol. *16*, 415–417.
7. Yang, X., Chen, A., PourNejatian, N., Shin, H.C., Smith, K.E., Parisien, C., Compas, C., Martin, C., Costa, A.B., Flores, M.G., et al. (2022). A large language model for electronic health records. NPJ Digit. Med. *5*, 194. https://doi.org/10.1038/s41746-022-00742-2.
8. (2023). Will ChatGPT transform healthcare? Nat. Med. *29*, 505–506. https://doi.org/10.1038/s41591-023-02289-5.
9. Gilbert, S., Harvey, H., Melvin, T., Vollebregt, E., and Wicks, P. (2023). Large language model AI chatbots require approval as medical devices. Nat. Med. *29*, 2396–2398. https://doi.org/10.1038/s41591-023-02412-6.
10. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., and Ting, D.S.W. (2023). Large language models in medicine. Nat. Med. *29*, 1930–1940. https://doi.org/10.1038/s41591-023-02448-8.
11. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. Nature *620*, 172–180. https://doi.org/10.1038/s41586-023-06291-2.
12. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief. Bioinform. *23*, bbac409. https://doi.org/10.1093/bib/bbac409.
13. Mann, D.L. (2023). Artificial Intelligence Discusses the Role of Artificial Intelligence in Translational Medicine: A JACC: Basic to Translational Science Interview With ChatGPT. JACC. Basic Transl. Sci. *8*, 221–223. https://doi.org/10.1016/j.jacbts.2023.01.001.
14. Uprety, D., Zhu, D., and West, H.J. (2023). ChatGPT-A promising generative AI tool and its implications for cancer care. Cancer *129*, 2284–2289. https://doi.org/10.1002/cncr.34827.
15. Agathokleous, E., Saitanis, C.J., Fang, C., and Yu, Z. (2023). Use of ChatGPT: What does it mean for biology and environmental science? Sci. Total Environ. *888*, 164154. https://doi.org/10.1016/j.scitotenv.2023.164154.
16. Li, S.W., Kemp, M.W., Logan, S.J.S., Dimri, P.S., Singh, N., Mattar, C.N.Z., Dashraath, P., Ramlal, H., Mahyuddin, A.P., Kanayan, S., et al. (2023). ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Am. J. Obstet. Gynecol. *229*, 172.e1–172.e12. https://doi.org/10.1016/j.ajog.2023.04.020.
17. 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. https://doi.org/10.1016/j.ebiom.2023.104512.
18. Kaneda, Y. (2023). In the Era of Prominent AI, What Role Will Physicians Be Expected to Play? QJM *116*, 881. https://doi.org/10.1093/qjmed/hcad099.
19. Galido, P.V., Butala, S., Chakerian, M., and Agustines, D. (2023). A Case Study Demonstrating Applications of ChatGPT in the Clinical Management of Treatment-Resistant Schizophrenia. Cureus *15*, e38166. https://doi.org/10.7759/cureus.38166.
20. Yeo, Y.H., Samaan, J.S., and Ng, W.H. (2023). The Application of GPT-4 in patient education and healthcare delivery. Clin. Mol. Hepatol. *29*, 821–822. https://doi.org/10.3350/cmh.2023.0183.
21. Zhavoronkov, A. (2023). Caution with AI-generated content in biomedicine. Nat. Med. *29*, 532. https://doi.org/10.1038/d41591-023-00014-w.
22. Li, R., Kumar, A., and Chen, J.H. (2023). How Chatbots and Large Language Model Artificial Intelligence Systems Will Reshape Modern Medicine: Fountain of Creativity or Pandora's Box? JAMA Intern. Med. *183*, 596–597. https://doi.org/10.1001/jamainternmed.2023.1835.
23. Ruksakulpiwat, S., Kumar, A., and Ajibade, A. (2023). Using ChatGPT in Medical Research: Current Status and Future Directions. J. Multidiscip. Healthc. *16*, 1513–1520. https://doi.org/10.2147/jmdh.S413470.
24. Blum, J., Menta, A.K., Zhao, X., Yang, V.B., Gouda, M.A., and Subbiah, V. (2023). Pearls and pitfalls of ChatGPT in medical oncology. Trends Cancer *9*, 788–790. https://doi.org/10.1016/j.trecan.2023.06.007.
25. Zhou, Z., Wang, X., Li, X., and Liao, L. (2023). Is ChatGPT an Evidence-based Doctor? Eur. Urol. *84*, 355–356. https://doi.org/10.1016/j.eururo.2023.03.037.
26. Perera Molligoda Arachchige, A.S. (2023). Large language models (LLM) and ChatGPT: a medical student perspective. Eur. J. Nucl. Med. Mol. Imag. *50*, 2248–2249. https://doi.org/10.1007/s00259-023-06227-y.
27. Munoz-Zuluaga, C., Zhao, Z., Wang, F., Greenblatt, M.B., and Yang, H.S. (2023). Assessing the Accuracy and Clinical Utility of ChatGPT in Laboratory Medicine. Clin. Chem. *69*, 939–940. https://doi.org/10.1093/clinchem/hvad058.
28. Liu, X., Wu, C., Lai, R., Lin, H., Xu, Y., Lin, Y., and Zhang, W. (2023). ChatGPT: when the artificial intelligence meets standardized patients in clinical training. J. Transl. Med. *21*, 447. https://doi.org/10.1186/s12967-023-04314-0.
29. Ayers, J.W., Zhu, Z., Poliak, A., Leas, E.C., Dredze, M., Hogarth, M., and Smith, D.M. (2023). Evaluating Artificial Intelligence Responses to Public Health Questions. JAMA Netw. Open *6*, e2317517. https://doi.org/10.1001/jamanetworkopen.2023.17517.
30. Sharma, P., and Parasa, S. (2023). ChatGPT and large language models in gastroenterology. Nat. Rev. Gastroenterol. Hepatol. *20*, 481–482. https://doi.org/10.1038/s41575-023-00799-8.
31. Jiang, L.Y., Liu, X.C., Nejatian, N.P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H.A., Laufer, I., Punjabi, P., et al. (2023). Health system-scale language models are all-purpose prediction engines. Nature *619*, 357–362. https://doi.org/10.1038/s41586-023-06160-y.
32. Thirunavukarasu, A.J. (2023). Large language models will not replace healthcare professionals: curbing popular fears and hype. J. R. Soc. Med. *116*, 181–182. https://doi.org/10.1177/01410768231173123.
33. Teixeira da Silva, J.A. (2023). Letter to the Editor in Response to article by Vaishya et al ChatGPT: Is this version good for healthcare and research. Diabetes Metab. Syndr. *17*, 102779. https://doi.org/10.1016/j.dsx.2023.102779.
34. Miloski, B. (2023). Opportunities for artificial intelligence in healthcare and *in vitro* fertilization. Fertil. Steril. *120*, 3–7. https://doi.org/10.1016/j.fertnstert.2023.05.006.
35. Korngiebel, D.M., and Mooney, S.D. (2021). Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digit. Med. *4*, 93. https://doi.org/10.1038/s41746-021-00464-x.
36. Peng, C., Yang, X., Chen, A., Smith, K.E., PourNejatian, N., Costa, A.B., Martin, C., Flores, M.G., Zhang, Y., Magoc, T., et al. (2023). A study of generative large language

model for medical research and healthcare. NPJ Digit. Med. *6*, 210. https://doi.org/10.1038/s41746-023-00958-w.

37. Thapa, S., and Adhikari, S. (2023). ChatGPT, Bard, and Large Language Models for Biomedical Research: Opportunities and Pitfalls. Ann. Biomed. Eng. *51*, 2647–2651. https://doi.org/10.1007/s10439-023-03284-0.
38. Bernstein, I.A., Zhang, Y.V., Govil, D., Majid, I., Chang, R.T., Sun, Y., Shue, A., Chou, J.C., Schehlein, E., Christopher, K.L., et al. (2023). Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. JAMA Netw. Open *6*, e2330320. https://doi.org/10.1001/jamanetworkopen.2023.30320.
39. Decker, H., Trang, K., Ramirez, J., Colley, A., Pierce, L., Coleman, M., Bongiovanni, T., Melton, G.B., and Wick, E. (2023). Large Language Model−Based Chatbot vs Surgeon-Generated Informed Consent Documentation for Common Procedures. JAMA Netw. Open *6*, e2336997. https://doi.org/10.1001/jamanetworkopen.2023.36997.
40. Wang, J., Huang, Y., Chen, C., Liu, Z., Wang, S., and Wang, Q. (2023). Software testing with large language model: Survey, landscape, and vision. Preprint at arXiv. https://doi.org/10.48550/arXiv.2307.07221.
41. Bowman, S.R. (2023). Eight Things to Know about Large Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2304.00612.
42. Gu, Y., Cao, J., Guo, Y., Qian, S., and Guan, W. (2023). Plan, Generate and Match: Scientific Workflow Recommendation with Large Language Models (Springer), pp. 86–102.
43. Lappin, S. (2023). Assessing the Strengths and Weaknesses of Large Language Models. J. Logic Lang. Inf. *33*, 9–20.
44. Arora, A., and Arora, A. (2023). The promise of large language models in health care. Lancet *401*, 641.
45. Nakaura, T., and Naganawa, S. (2023). Writing medical papers using large-scale language models: a perspective from the Japanese Journal of Radiology. Jpn. J. Radiol. *41*, 457–458.
46. Arighi, C., Brenner, S., and Lu, Z. (2023). LARGE LANGUAGE MODELS (LLMS) AND CHATGPT FOR BIOMEDICINE (World Scientific), pp. 641–644.
47. Cascella, M., Montomoli, J., Bellini, V., and Bignami, E. (2023). Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J. Med. Syst. *47*, 33.
48. Lower, K., Seth, I., Lim, B., and Seth, N. (2023). ChatGPT-4: transforming medical education and addressing clinical exposure challenges in the post-pandemic era. Indian J. Orthop. *57*, 1527–1544.
49. Zhuang, Y., Yu, Y., Wang, K., Sun, H., and Zhang, C. (2023). ToolQA: A Dataset for LLM Question Answering with External Tools. Preprint at arXiv. https://doi.org/10.48550/arXiv.2306.13304.
50. Robinson, J., Rytting, C.M., and Wingate, D. (2022). Leveraging Large Language Models for Multiple Choice Question Answering. Preprint at arXiv. https://doi.org/10.48550/arXiv.2210.12353.
51. Extance, A. (2023). ChatGPT has entered the classroom: how LLMs could transform education. Nature *623*, 474–477.
52. Moore, S., Tong, R., Singh, A., Liu, Z., Hu, X., Lu, Y., Liang, J., Cao, C., Khosravi, H., and Denny, P. (2023). Empowering Education with LIms-The Next-Gen Interface and Content Generation (Springer), pp. 32–37.
53. Dave, T., Athaluri, S.A., and Singh, S. (2023). ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front. Artif. Intell. *6*, 1169595.
54. Dunn, A., Dagdelen, J., Walker, N., Lee, S., Rosen, A.S., Ceder, G., Persson, K., and Jain, A. (2022). Structured Information Extraction from Complex Scientific Text with Fine-Tuned Large Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2212.05238.
55. Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., and Hu, X. (2023). Harnessing the Power of LIms in Practice: A Survey on Chatgpt and beyond. Preprint at arXiv. https://doi.org/10.48550/arXiv.2304.13712.
56. Tu, X., Zou, J., Su, W.J., and Zhang, L. (2023). What Should Data Science Education Do with Large Language Models?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2307.02792.
57. Dias, R., and Torkamani, A. (2019). Artificial intelligence in clinical and genomic diagnostics. Genome Med. *11*, 70.
58. Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A.I., Almohareb, S.N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H.A., et al. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med. Educ. *23*, 689.
59. Han, C., Kim, D.W., Kim, S., You, S.C., Park, J.Y., Bae, S., and Yoon, D. (2024). Evaluation of GPT-4 for 10-year cardiovascular risk prediction: insights from the UK Biobank and KoGES data. iScience *27*, 109022.
60. Benary, M., Wang, X.D., Schmidt, M., Soll, D., Hilfenhaus, G., Nassir, M., Sigler, C., Knödler, M., Keller, U., Beule, D., et al. (2023). Leveraging large language models for decision support in personalized oncology. JAMA Netw. Open *6*, e2343689.
61. Harris, E. (2023). Large language models answer medical questions accurately, but can't match clinicians' knowledge. JAMA *330*, 792–794.
62. Liu, J., Zheng, J., Cai, X., Wu, D., and Yin, C. (2023). A descriptive study based on the comparison of ChatGPT and evidence-based neurosurgeons. iScience *26*, 107590.
63. Shah, N.H., Entwistle, D., and Pfeffer, M.A. (2023). Creation and adoption of large language models in medicine. JAMA *330*, 866–869.
64. Zhang, T., and Feng, T. (2023). Application and technology of an open source AI large language model in the medical field. Radiol. Sci. *2*, 96–104.
65. Omiye, J.A., Lester, J.C., Spichak, S., Rotemberg, V., and Daneshjou, R. (2023). Large language models propagate race-based medicine. NPJ Digit. Med. *6*, 195.
66. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., and Ting, D.S.W. (2023). Large language models in medicine. Nat. Med. *29*, 1930–1940.
67. Zhang, P., and Kamel Boulos, M.N. (2023). Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges. Future Internet *15*, 286.
68. Nazi, Z.A., and Peng, W. (2023). Large Language Models in Healthcare and Medical Domain: A Review. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.06775.
69. Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., and Xie, X. (2023). Large Language Models Understand and Can Be Enhanced by Emotional Stimuli. Preprint at arXiv. https://doi.org/10.48550/arXiv.2307.11760.
70. Spinewine, A., Evrard, P., and Hughes, C. (2021). Interventions to optimize medication use in nursing homes: a narrative review. Eur. Geriatr. Med. *12*, 551–567.
71. Eisenstein, E., Kopacek, C., Cavalcante, S.S., Neves, A.C., Fraga, G.P., and Messina, L.A. (2020). Telemedicine: a Bridge Over Knowledge Gaps in Healthcare. Curr. Pediatr. Rep. *8*, 93–98. https://doi.org/10.1007/s40124-020-00221-w.
72. Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G., and Klang, E. (2023). Large language models (llms) and empathy-a systematic review. Preprint at medRxiv. https://doi.org/10.1101/2023.08.07.23293769.
73. Zheng, Z., Liao, L., Deng, Y., and Nie, L. (2023). Building Emotional Support Chatbots in the Era of LIms. Preprint at arXiv. https://doi.org/10.48550/arXiv.2308.11584.
74. Qureshi, R., Shaughnessy, D., Gill, K.A.R., Robinson, K.A., Li, T., and Agai, E. (2023). Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? Syst. Rev. *12*, 72. https://doi.org/10.1186/s13643-023-02243-z.
75. Chen, Q., Du, J., Hu, Y., Keloth, V.K., Peng, X., Raja, K., Zhang, R., Lu, Z., and Xu, H. (2023). Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.16326.
76. Atas Guvenilir, H., and Doğan, T. (2023). How to approach machine learning-based prediction of drug/compound–target interactions. J. Cheminform. *15*, 16.
77. Toufiq, M., Rinchai, D., Bettacchioli, E., Kabeer, B.S.A., Khan, T., Subba, B., White, O., Yurieva, M., George, J., Jourde-Chiche, N., et al. (2023). Harnessing large language models (LLMs) for candidate gene prioritization and selection. J. Transl. Med. *21*, 728.
78. Hegedűs, T., Geisler, M., Lukács, G.L., and Farkas, B. (2022). Ins and outs of AlphaFold2 transmembrane protein structure predictions. Cell. Mol. Life Sci. *79*, 73.
79. Valavanidis, A. AlphaFold Protein Structure Database Predicted Millions of 3D Structures.
80. Wilson, A.E., Lehmann, C.U., Saleh, S.N., Hanna, J., and Medford, R.J. (2021). Social media: a new tool for outbreak surveillance. Antimicrob. Steward. Healthc. Epidemiol. *1*, e50.
81. Aiello, A.E., Renson, A., and Zivich, P.N. (2020). Social media-and internet-based disease surveillance for public health. Annu. Rev. Public Health *41*, 101–118.
82. Ueda, D., Walston, S.L., Matsumoto, T., Deguchi, R., Tatekawa, H., and Miki, Y. (2023). Evaluating GPT-4-Based ChatGPT's Clinical Potential on the NEJM Quiz.
83. De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G.P., Ferragina, P., Tozzi, A.E., and Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front. Public Health *11*, 1166120. https://doi.org/10.3389/fpubh.2023.1166120.

84. Ali, S.R., Dobbs, T.D., Hutchings, H.A., and Whitaker, I.S. (2023). Using ChatGPT to write patient clinic letters. Lancet Digit. Health *5*, e179–e181. https://doi.org/10.1016/s2589-7500(23)00048-1.
85. Haruna-Cooper, L., and Rashid, M.A. (2023). GPT-4: the future of artificial intelligence in medical school assessments. J. R. Soc. Med. *116*, 218–219. https://doi.org/10.1177/01410768231181251.
86. Misal, D., and Misal, D. (2020). Indian Startups Revolutionizing the Healthcare Sector with AI.
87. Sezgin, E. (2023). Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. Digit. Health *9*, 20552076231186520. https://doi.org/10.1177/20552076231186520.
88. Tursunbayeva, A., and Renkema, M. (2022). Artificial intelligence in health-care: implications for the job design of healthcare professionals. Asia Pac. J. Hum. Resour. *61*, 845–887.
89. Kwok, K.O., Wei, W.I., Tsoi, M.T.F., Tang, A., Chan, M.W.H., Ip, M., Li, K.-K., and Wong, S.Y.S. (2023). How can we transform travel medicine by leveraging on AI-powered search engines? J. Travel Med. *30*, taad058. https://doi.org/10.1093/jtm/taad058.
90. Cheng, K., Wu, H., and Li, C. (2023). ChatGPT/GPT-4: enabling a new era of surgical oncology. Int. J. Surg. *109*, 2549–2550. https://doi.org/10.1097/js9.0000000000000451.
91. Cheng, K., Wu, C., Gu, S., Lu, Y., Wu, H., and Li, C. (2023). WHO declares end of COVID-19 global health emergency: lessons and recommendations from the perspective of ChatGPT/GPT-4. Int. J. Surg. *109*, 2859–2862. https://doi.org/10.1097/js9.0000000000000521.
92. Lu, Y., Qi, S., Cheng, K., and Wu, H. (2023). WHO declares end of mpox global health emergency: first glance from a perspective of ChatGPT/GPT-4. Int. J. Surg. *109*, 3217–3218. https://doi.org/10.1097/js9.0000000000000543.
93. Kanjee, Z., Crowe, B., and Rodman, A. (2023). Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA *330*, 78–80. https://doi.org/10.1001/jama.2023.8288.
94. Preiksaitis, C., Sinsky, C.A., and Rose, C. (2023). ChatGPT is not the solution to physicians' documentation burden. Nat. Med. *29*, 1296–1297. https://doi.org/10.1038/s41591-023-02341-4.
95. Komorowski, M., Del Pilar Arias López, M., and Chang, A.C. (2023). How could ChatGPT impact my practice as an intensivist? An overview of potential applications, risks and limitations. Intensive Care Med. *49*, 844–847. https://doi.org/10.1007/s00134-023-07096-7.
96. Kiros, R., Salakhutdinov, R., and Zemel, R. (2014). Multimodal Neural Language Models (PMLR), pp. 595–603.
97. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., and Yu, T. (2023). Palm-e: An embodied multimodal language model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.03378.
98. Zhang, T., Liu, N., Xu, J., Liu, Z., Zhou, Y., Yang, Y., Li, S., Huang, Y., and Jiang, S. (2023). Flexible electronics for cardiovascular healthcare monitoring. Innovation *4*, 100485. https://doi.org/10.1016/j.xinn.2023.100485.
99. Volpe, N.J., and Mirza, R.G. (2023). Chatbots, Artificial Intelligence, and the Future of Scientific Reporting. JAMA Ophthalmol. *141*, 824–825. https://doi.org/10.1001/jamaophthalmol.2023.3344.
100. Raita, Y., Goto, T., Faridi, M.K., Brown, D.F.M., Camargo, C.A., Jr., and Hasegawa, K. (2019). Emergency department triage prediction of clinical outcomes using machine learning models. Crit. Care *23*, 64. https://doi.org/10.1186/s13054-019-2351-7.
101. Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., and Mian, A.S. (2023). A Comprehensive Overview of Large Language Models. Preprint at ArXiv. https://doi.org/10.48550/arXiv.2307.06435.
102. Hardy, M., Sucholutsky, I., Thompson, B., and Griffiths, T. (2023). Large Language Models Meet Cognitive Science: LLMs as Tools, Models, and Participants, p. 45.
103. Glukhov, D., Shumailov, I., Gal, Y., Papernot, N., and Papyan, V. (2023). LLM Censorship: A Machine Learning Challenge or a Computer Security Problem?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2307.10719.
104. Kumar, V., Gleyzer, L., Kahana, A., Shukla, K., and Karniadakis, G.E. (2023). Mycrunchgpt: A llm assisted framework for scientific machine learning. J. Mach. Learn. Model. Comput. *4*, 41–72.
105. Ali, T., and Kostakos, P. (2023). Huntgpt: Integrating Machine Learning-Based Anomaly Detection and Explainable Ai with Large Language Models (LLMs). Preprint at arXiv. https://doi.org/10.48550/arXiv.2309.16021.
106. Zhang, L., Zhang, Y., Ren, K., Li, D., and Yang, Y. (2023). MLCopilot: Unleashing the Power of Large Language Models in Solving Machine Learning Tasks. Preprint at arXiv. https://doi.org/10.48550/arXiv.2304.14979.
107. Daneshjou, R., Smith, M.P., Sun, M.D., Rotemberg, V., and Zou, J. (2021). Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. JAMA Dermatol. *157*, 1362–1369.
108. González-Sendino, R., Serrano, E., and Bajo, J. (2024). Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making. Future Generat. Comput. Syst. *155*, 384–401.
109. Ward, E., and Gross, C. (2023). Evolving Methods to Assess Chatbot Performance in Health Sciences Research. JAMA Intern. Med. *183*, 1030–1031. https://doi.org/10.1001/jamainternmed.2023.2567.
110. Butte, A.J. (2023). Artificial Intelligence-From Starting Pilots to Scalable Privilege. JAMA Oncol. *9*, 1341–1342. https://doi.org/10.1001/jamaoncol.2023.2867.
111. Hu, Z.-Y., Han, F.-J., Yu, L., Jiang, Y., and Cai, G. (2023). AI-link omnipotent pathological robot: Bridging medical meta-universe to real-world diagnosis and therapy. Innovation *4*, 100494. https://doi.org/10.1016/j.xinn.2023.100494.
112. Ahmadhil, A. (2023). AMA Issues New Principles for Use of AI in Medicine (Infectious Disease Advisor).
113. Levac, D., Colquhoun, H., and O'Brien, K.K. (2010). Scoping studies: advancing the methodology. Implement. Sci. *5*, 69. https://doi.org/10.1186/1748-5908-5-69.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
R version 4.0.3	R software	Description: A software environment for statistical computing and graphics. Provider: R Core Team Obtained From: The R software can be freely downloaded from the Comprehensive R Archive Network (CRAN) at https://cran.r-project.org.
ArcGIS, version 10.0	ArcGIS Software	**Description**: A geographic information system for working with maps and geographic information. **Manufacturer**: Environmental Systems Research Institute (ESRI) **Obtained From**: The ArcGIS software is available for purchase or through a licensing agreement from ESRI. More information can be found on their official website at https://www.esri.com.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Yi-Da Tang (e-mail: tangyida@bjmu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The titles and DOI information of the literature summarized in this study are provided as attachments.
- This paper does not report original code.
- Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Study design

This research was based on a scoping review methodology, adopting Arksey and O'Malley's structured five-step approach,[113] which encompasses (1) delineating the research query, (2) pinpointing pertinent studies, (3) selecting studies, (4) extracting and charting data, and (5) collating, synthesising, and presenting the results. The study was registered on **OSF REGISTRIES** and can be accessed at https://osf.io/p7cek.

Literature search

Our research involved comprehensive database searches across PubMed, Embase, Web of Science, and Scopus until 20 July 2023. The search was conducted using the following keywords: (1) 'Large Language Model' or 'LLM'; (2) 'ChatGPT' or 'GPT' or 'GPT-3' or 'GPT-4'; and (3) 'Generative Pre-trained Transformer'. Given the study's emphasis on LLMs' applications in the medical field, only articles within the 'medicine' category from the Scopus database were included. As these databases collectively encompass articles from preprint repositories, such as arXiv, bioRxiv, medRxiv, and Research Square, the search strategy effectively captured cutting-edge research.

Study selection

The initial steps involved removing redundant articles. A two-phase selection process followed, with the first phase consisting of a cursory review of titles, abstracts, and keywords. The criteria for initial inclusion were as follows: (1) study relevance to LLMs, (2) alignment with the medical field, and (3) articles written in either English or Chinese. The subsequent selections underwent more rigorous evaluations and relied on full-text reviews. Notably, the article type was not a determinant of inclusion. This dual-phase selection process was carried out independently by two reviewers (DL and XM). Disparities in choices were reconciled through discussions, and if a consensus remained elusive, a third reviewer (XY) arbitrated the final decision.

Data extraction

Data extraction was performed by eight reviewers (XM, DL, ZG, YL, XC, CQ, TG, and SM) using a bespoke extraction form. The extracted data included article title, authors and their countries, journal name, publication date, and the main findings of the article summarized by the reviewers. For each article, two reviewers were arranged to extract data respectively and reconcile their respective data records after extraction. In scenarios where a consensus could not be reached, a third reviewer (XY) was consulted to make the final decision.

Data integration

Based on the main findings of the articles, reviewers further collated the types of articles, organ and systems affected by the diseases (if clearly indicated in the articles), and the role of LLMs in transforming medical practices. In this study, article types were classified into original study and non-original study based on whether the study provided authors' own data or results through independent investigation or experiment. The original studies were always often referred to in journals as original research or research letter. While, the non-original studies were review, commentary, editorial, and other forms in the journals. The organ and systems affected by the diseases were classified into nervous, endocrine, respiratory, genitourinary, musculoskeletal, digestive, immune, circulatory, eye, ear-nose-throat, oral and maxillofacial organ and systems. In addition, during the data extraction process, we found that several studies focused on mental and psychological diseases. To cover the diseases more comprehensively, psychiatry was included a separate category, which would not overlap with other organ systems. The roles of LLMs in transforming medical practices included medical writing, education and training, project management, diagnosis and treatment, patient communication, nursing, and other areas.

A chronological trend analysis of publications was conducted by monthly tabulation, which was visually represented using stacked bar charts. Furthermore, distinctions were made between original studies and non-original studies. The geographical distribution across varying time frames was depicted using maps. A categorical breakdown highlighted research across diverse organ systems. The applications of LLMs in transforming medical practices were integrated and described in details classified by different roles.

Data description and statistics by different categories were performed using R version 4.0.3 (R Core Team). Mapping was performed with ArcGIS, version 10.0 (ESIR).