Biomedical Journal 48 (2025) 100868

Contents lists available at ScienceDirect

Biomedical Journal

journal homepage: www.elsevier.com/locate/bj

Review Article

Roles and potential of Large language models in healthcare: A comprehensive review

Chihung Lin [a,d], Chang-Fu Kuo [a,b,c,*]

[a] Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan
[b] Division of Rheumatology, Allergy, and Immunology, Chang Gung Memorial Hospital, Taoyuan, Taiwan
[c] Division of Rheumatology, Orthopaedics and Dermatology, School of Medicine, University of Nottingham, Nottingham, UK
[d] Department of Artificial Intelligence, Chang Gung University, Taoyuan, Taiwan

ARTICLE INFO

Keywords:
Large language models
Healthcare
Artificial intelligence
Clinical decision support
Patient communication

ABSTRACT

Large Language Models (LLMs) are capable of transforming healthcare by demonstrating remarkable capabilities in language understanding and generation. They have matched or surpassed human performance in standardized medical examinations and assisted in diagnostics across specialties like dermatology, radiology, and ophthalmology. LLMs can enhance patient education by providing accurate, readable, and empathetic responses, and they can streamline clinical workflows through efficient information extraction from unstructured data such as clinical notes. Integrating LLM into clinical practice involves user interface design, clinician training, and effective collaboration between Artificial Intelligence (AI) systems and healthcare professionals. Users must possess a solid understanding of generative AI and domain knowledge to assess the generated content critically. Ethical considerations to ensure patient privacy, data security, mitigating biases, and maintaining transparency are critical for responsible deployment. Future directions for LLMs in healthcare include interdisciplinary collaboration, developing new benchmarks that incorporate safety and ethical measures, advancing multimodal LLMs that integrate text and imaging data, creating LLM-based medical agents capable of complex decision-making, addressing underrepresented specialties like rare diseases, and integrating LLMs with robotic systems to enhance precision in procedures. Emphasizing patient safety, ethical integrity, and human-centered implementation is essential for maximizing the benefits of LLMs, while mitigating potential risks, thereby helping to ensure that these AI tools enhance rather than replace human expertise and compassion in healthcare.

1. Introduction

Artificial intelligence (AI) encompasses methods and systems designed to perform tasks that typically require human intelligence. Within AI, machine learning (ML) refers to algorithms that automatically learn patterns from data, while natural language processing (NLP) focuses on enabling computers to interpret, generate, and utilize human language [1,2]. Large language models (LLMs)—exemplified by ChatGPT, Claude, and Gemini—leverage recent breakthroughs in deep learning and massive datasets to achieve impressive fluency and contextual understanding, and contributed significantly to medicine, education, and research by leveraging deep learning techniques and massive datasets [3–5]. In addition to commercial LLM, specialized LLMs are emerging for specific tasks, such as Med-PaLM for clinical question answering and BioBERT for biomedical text mining [6]. Some of these models focus on narrower specialties (e.g., radiology, oncology), leveraging domain-specific corpora to refine performance. This diversity underscores how LLMs can be tailored to various healthcare domains, from summarizing clinical trial literature to assisting in mental health triage.

Potential applications of LLMs span multiple healthcare domains, including real-time patient interaction through medical chatbots for support, triage, and education; enhanced clinical decision support by assisting with diagnoses, treatment recommendations, and drug information; automated or semi-automated imaging and reporting, particularly in radiology, pathology, and nuclear medicine; and data extraction that summarizes large volumes of electronic health records (EHRs) and research literature for more rapid insight [7]. LLMs excel in these areas

* Corresponding author. Division of Rheumatology, Allergy, and Immunology, Chang Gung Memorial Hospital, Address: No.5, Fuxing St., Guishan Dist., Taoyuan City, Taiwan.
E-mail address: zandis@gmail.com (C.-F. Kuo).

https://doi.org/10.1016/j.bj.2025.100868
Received 20 November 2024; Received in revised form 14 April 2025; Accepted 28 April 2025
Available online 29 April 2025

due to their ability to handle unstructured textual data, synthesize complex information, and generate coherent, context-aware responses [8]. Their flexible language capabilities position them uniquely for scenarios requiring reliable communication and meticulous documentation, both of which are foundational to effective healthcare delivery. However, integrating such powerful AI tools into healthcare settings raises essential questions about their accuracy, reliability, ethics, and the changing nature of medical practice [9]. This review synthesizes current research on LLMs' applications, performance, and limitations in healthcare settings. By examining studies across various medical specialties and uses, we aimed to provide healthcare professionals with a thorough understanding of the current state of LLMs in medicine, their potential advantages, and the challenges that must be addressed for successful deployment.

1.1. Overview of large language models

LLMs are sophisticated systems built using deep learning techniques, primarily leveraging the 'Transformer' architecture to process sequential data efficiently. These models undergo large-scale pre-training using vast datasets to capture semantic and syntactic patterns. The concept of scaling laws is crucial to their performance, as it highlights how increasing both model size and data can lead to significant gains in understanding and generating human-like language [10]. The capabilities of LLMs are fundamentally rooted in three key elements: advanced language modeling to predict and generate meaningful text sequences, the Transformer architecture for managing long-range dependencies in text, and extensive pre-training, which equips the models with broad contextual knowledge applicable across domains [11].

Language models are probabilistic models that predict the likelihood of a sequence of tokens (words, subwords, or morphemes) [8]. They estimate the probability of a sequence based on previous tokens, making them effective at predicting the next token or generating new sequences. This capability is central to understanding and generating language in LLMs, and the transformer architecture is at the core of modern LLMs [12]. Unlike recurrent neural networks (RNNs), which process tokens sequentially and struggle with long-range dependencies, the Transformer uses self-attention mechanisms. This approach allows the efficient modeling of long texts and enables parallel processing, making it highly scalable. Scalability is crucial for training large models on massive datasets, leading to improved performance [13].

Large-scale pre-training involves training LLMs on extensive corpora, such as Wikipedia, Common Crawl, and Books, using masked language modeling (MLM) and next-token prediction (NTP). In MLM, a portion of the input text is masked, and the model is trained to predict the masked tokens, capturing the semantic and syntactic relationships among tokens. NTP focuses on predicting the next token in a sequence based on previous tokens, helping the model generate coherent sequences. Scaling laws suggest that increasing the model and data sizes proportionally can lead to substantial performance gains [10]. This scaling is critical for achieving state-of-the-art results in various natural language processing (NLP) tasks, as seen in GPT-3 and PaLM models [14].

1.2. Evolution of medical large language models

The evolution of medical LLMs has been marked by significant advances in model architecture, training data, and fine-tuning techniques tailored explicitly for healthcare applications. This progression includes:

1. Early Biomedical Language Models: Initial attempts focused on pre-training existing architectures like BERT on biomedical corpora. Models such as BioBERT [15] and ClinicalBERT [16] have improved performance on tasks like named entity recognition and relation extraction in biomedical texts.
2. Specialized Medical LLMs: Researchers have started developing LLMs specifically designed for medical applications. Models like MEDITRON have been trained from scratch on vast amounts of medical literature and electronic health records, incorporating domain-specific knowledge directly into their pre-training process [17].
3. Multimodal Medical LLMs: Recent advances have led to the development of multimodal LLMs capable of processing text and medical imaging data. Models such as Med-PaLM have shown promising results in tasks like medical visual question answering and automated diagnosis based on clinical images and notes [8]. The latest developments focus on optimizing existing large models for medical tasks through advanced fine-tuning and prompting techniques. By leveraging these approaches, models like MedGemini have remarkably performed medical licensing examinations and clinical reasoning tasks [18].

[Fig. 1] outlines LLMs that are applied or specialized in the medical field. Currently, the primary focus of development and testing revolves around GPT and LLaMA. The diagram also indicates that since the release of ChatGPT in November 2022, there have been new opportunities for applications in the medical field. TAME is a model jointly developed by six major industries in Taiwan, with the healthcare industry being one of its primary sectors. Consequently, throughout 2023 and 2024, numerous medical-related LLM models have been developed using clinical data or medical research papers. These advancements suggest that future applications will better meet the specific needs of the medical domain.

1.3. Applications of medical large language models (LLMs) in healthcare

[Table 1] highlights potential and diverse applications of medical LLMs in healthcare, including their roles in medical examinations (e.g., ChatGPT passing the USMLE), clinical decision-making (e.g., SkinGPT-4 for dermatology), patient communication (e.g., multilingual support for autistic patients), and information extraction from health records. These LLMs can enhance medical education, diagnostics, patient engagement, and research, illustrating their potential impact on healthcare practices.

1.4. Performance on medical examinations and clinical reasoning

LLMs have demonstrated impressive abilities in standardized medical examinations, often matching or surpassing human performance. For example, Kung et al. reported that ChatGPT passed the United States Medical Licensing Examination (USMLE) with a score above the passing threshold, suggesting that LLMs possess medical knowledge comparable to medical graduates [19]. Similarly, Tsoutsanis found that Bing Chat outperformed other LLMs, including Llama 2 and ChatGPT-3.5, on the Multi-Specialty Recruitment Assessment (MSRA), even surpassing some human participants [20]. LLMs have also excelled in specialty-specific examinations; Nakajima et al. showed that GPT-4 surpassed its predecessor in the Japanese Orthopaedic Association Board Exam, demonstrating its advanced understanding of orthopedic knowledge [21], while Park et al. evaluated dermatology board exam performance and highlighted the variability in results depending on question complexity, emphasizing the need for specialty-specific assessments [22].

1.5. Clinical reasoning and decision-making

Beyond standardized examinations, the capabilities of LLMs in clinical reasoning tasks have also been assessed, demonstrating promising results in real-world medical scenarios in clinical decision support and diagnosis. [Fig. 2] illustrates the integration of LLM into clinical decision support workflows. The workflow begins by aggregating healthcare data sources—including electronic health records, medical imaging, and structured clinical data—alongside patient-specific clinical contexts. These inputs serve as the foundation for LLM analysis. The central LLM processing pipeline performs sophisticated operations including natural

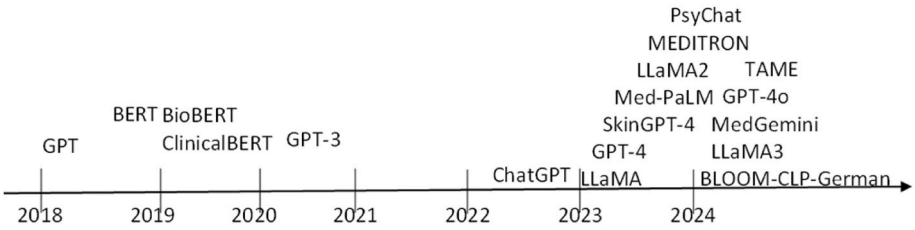

Fig. 1. A history of the development and evolution of Med-LLM.

Table 1
Applications of medical large language models (LLMs) in healthcare.

Application Area	Description	Examples of LLM Use	Challenges
Medical Examinations and Clinical Reasoning	LLMs are used for clinical reasoning tasks, as demonstrated by passing clinical board examinations.	- ChatGPT passed the USMLE [19, 71] - Bing Chat outperforms others on the Multi-specialty Recruitment Assessment (MSRA) [20]	- Variable performance across different specialties - Need for comprehensive evaluation before -clinical implementation
Clinical Decision Support and Diagnosis	LLMs assist in clinical decision-making and diagnosis across various specialties.	- SkinGPT-4 for dermatological diagnosis [27] - GPT-4 assessed emergency department acuity [23]	- Need for human oversight - Potential for errors in complex cases - Variability in performance across tasks
Patient Communication and Education	LLMs provide accurate and empathetic responses to patient queries, enhancing patient education and engagement.	- LLM-based AI chat for improving heart failure education [33] - PsyChat for mental health support [37] - Multilingual support for autistic patients [35]	- Readability levels often exceed patient comprehension [35] - Ensuring consistency with medical guidelines - Need for validation
Information Extraction and Medical Record Analysis	LLMs extract relevant information from electronic health records and medical literature, aiding research and clinical workflows.	- Extracting data from breast cancer pathology reports [40] - Summarizing radiology reports for glioblastoma patients [42]	- Need for careful validation - Dealing with sensitive medical information - Data privacy concerns

language processing, multimodal data integration, targeted knowledge retrieval, and clinical reasoning capabilities. The workflow incorporates specialized LLM-based medical agents that autonomously execute defined clinical tasks, which serves as a transition from passive decision support toward active workflow participation. The system generates diagnostic suggestions, treatment recommendations, and documentation while maintaining healthcare provider decision authority. Such a design addresses concerns identified in comparative studies regarding performance variability and clinical accuracy. A critical feedback loop from patient outcomes facilitates continuous model refinement, particularly addressing limitations identified in specialty applications and rare disease diagnostics. This iterative approach ensures ongoing enhancement of LLM capabilities and maintains a careful balance between artificial intelligence functionalities and human clinical expertise. Effective implementation thus recognizes both technological challenges and necessary organizational adaptations.

Recent studies underscore the diverse potential of LLMs. Williams et al. evaluated the ability of GPT-4 to assess clinical acuity in emergency department settings. They found that GPT-4 accurately identified high-acuity patients, performing comparably with human physician reviewers, thus highlighting its potential role in emergency triage [23]. Young et al. investigated the diagnostic accuracy of a specialized LLM for rare pediatric disease cases, although improvements are needed due to existing limitations [24]. Comparisons with human medical professionals have had mixed outcomes. Ye et al. compared the responses of ChatGPT with those of rheumatologists on patient-generated questions. Although patients rated LLM-generated and physician responses similarly, rheumatologists considered the AI-generated answers to be lower quality [25]. This discrepancy points to a gap between patient satisfaction and clinical accuracy, highlighting the need for expert evaluation when validating AI tools for healthcare applications.

Specialized contexts further illustrate LLM utility and variability. Shin et al. demonstrated that GPT-4 effectively provided drug information and guidance in community pharmacy settings, indicating its utility in pharmacy practice [26]. Specialty-specific applications have also shown promise; SkinGPT-4 for dermatological diagnosis performed comparably with board-certified dermatologists [27], while a radiology-specific LLM generated accurate impressions from radiology reports [28], and a specialized Chinese ophthalmology LLM performed well on board examinations and diagnostic tasks [29]. The ability of LLMs to handle rare complex cases has also been explored, with Young et al. showing that a custom model had mixed results in diagnosing rare pediatric diseases, underlining its potential and limitations [24].

Integrating structured medical knowledge to enhance LLM diagnostic capabilities represents a critical development pathway. Gao et al. explored leveraging a medical knowledge graph into LLMs for diagnosis prediction, demonstrating improved performance and interpretability [30]. Direct comparisons with human clinicians revealed that some advanced LLMs outperformed oncology residents and fellows on specific colorectal cancer-related queries [31].

While LLMs have shown capabilities that match or even surpass human experts in certain aspects, such as emergency triage and oncology-related queries, the variability in their performance—especially in specialized and complex cases—highlights the need for further refinement and expert validation. Effective collaboration between clinicians and LLMs, supported by structured continuous feedback mechanisms, promises enhanced patient outcomes and optimized healthcare workflows. Overall, LLMs have the potential to augment clinical practice significantly, but ensuring their safety, reliability, and integration with human workflows is essential for maximizing their impact in real-world medical applications.

1.6. Patient communication and education

LLMs are also used to enhance patient communication and education. Studies have shown that LLMs can generate accurate readable responses to patient questions on various medical topics. One study assessed the accuracy and readability of material on kidney stones produced by LLMs and compared this to official content from recognized urologic organizations [32]. This revealed that LLMs generally produced accurate readable information; however, there were some gaps in data

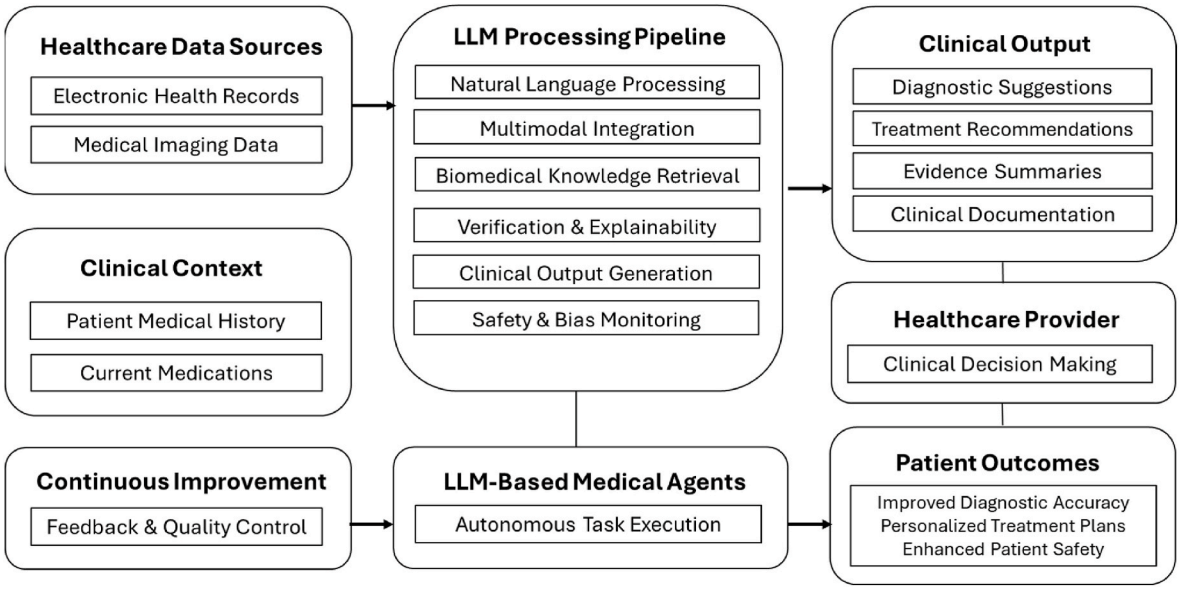

Fig. 2. Workflow Diagram Showing LLM Integration in the Clinical Decision Support process.

completeness. Another evaluation focused on the ability of LLMs to provide accurate consistent responses to patient questions about heart failure; generally, the LLMs delivered satisfactory information [33]. While LLMs can generate empathetic informative responses, concerns about health literacy remain. For instance, readability analysis of LLM responses to patient messages found that the generated text often exceeded the recommended readability levels for effective communication [34].

The ability of LLMs to operate across multiple languages has potential for enhancing healthcare communication in diverse populations. This ability is crucial in multicultural societies where healthcare systems must serve diverse populations with varying linguistic needs. One study specifically evaluated LLMs in the context of answering health-related questions for autistic patients in Chinese, demonstrating that LLMs could effectively provide culturally and linguistically tailored health information [35]. Studies emphasize the need for effective strategies to overcome language barriers, to enhance patient safety and communication quality [36]. The multilingual medical language model MMedC advances this by enhancing access to medical information for diverse linguistic groups [37]. By integrating multilingual data, LLMs to bridge communication gaps in multicultural and multilingual settings, thereby improving patient access to healthcare information.

Specialty-specific applications in patient education, such as for atopic dermatitis, have shown promise in providing information to patients that could enhance disease management and overall satisfaction [38]. LLMs have also been explored in mental health support, with systems like PsyChat showing the potential to expand access to mental health resources through AI-driven communication tools [37]. However, implementing such tools must consider ethical aspects, including maintaining patient privacy and ensuring the accuracy of therapeutic advice [39].

1.7. *Information extraction and medical record analysis*

The capacity of LLMs to process and analyze vast volumes of unstructured text has opened new avenues in healthcare, particularly in information extraction from medical records and scientific literature. These capabilities make LLMs valuable for generating meaningful insights from clinical data, supporting decision-making processes, and reducing manual labor in healthcare environments. The versatility of LLMs allows them to transition seamlessly between different healthcare domains—from analyzing electronic health records (EHRs) to synthesizing findings from scientific literature—providing an integrated approach to improving clinical workflows, diagnostics, and medical research.

A comparative study by Luo et al. used LLM-based zero-shot inference for extracting key information from breast cancer pathology reports and compared it to task-specific supervised classification [40]. Their findings indicated that LLMs could significantly reduce the burden of data labeling. Williams et al. evaluated GPT-4 for assessing clinical acuity in emergency departments based on EHR data and demonstrated that the model could accurately identify patients with higher acuity levels [23]. This suggests that LLMs have great potential for triage and resource allocation. The analysis of radiology reports is another area where LLMs have demonstrated substantial utility. Yasaka et al. developed a fine-tuned LLM to extract information on patients undergoing pretreatment for lung cancer using data from a Picture Archiving and Communication System (PACS) [41]. This model had high accuracy and efficiency compared to traditional manual review, suggesting that LLMs could be instrumental in streamlining oncology workflows and reducing clinician workload. Furthermore, Laukamp et al. used GPT-4 to summarize radiology reports for glioblastoma patients, effectively extracting and synthesizing essential information to create structured summaries [42].

1.8. *Integration into clinical workflows*

Integrating LLMs into clinical workflows requires addressing critical factors, including user interface design, training, and education, as well as effective collaboration between AI and healthcare professionals. Findings indicated that LLMs, like the BLOOM-CLP-German model, could generate useable medical reports with high accuracy, but their deployment requires adherence to stringent privacy standards and careful model alignment with clinical data. While the results were promising, the study emphasized the need for refinement in model training and integration before widespread implementation [43]. The integration of AI in healthcare requires a human-centered approach focused on designing interfaces that effectively enhance human-machine interaction to support diagnostic and decision-making processes [44]. Beyond explainability, successful AI adoption in clinical settings demands transparent interactions to ensure trust and usability within complex socio-technical systems [45,46].

Training and education are essential for effectively leveraging LLMs in healthcare. Future clinical educators must possess comprehensive

clinical competence and further develop empathy, the ability to collaborate with AI, and medical literacy to critically critique and evaluate the generated information. LLM-driven chatbots can significantly contribute to medical education by providing tailored, interactive content that enhances learning experiences, particularly in specialized fields like neuro-ophthalmology [47]. LLMs can deliver accurate medical management plans, summarize recent literature, and generate visual aids to help understand complex conditions, making them valuable for specialized medical education. A systematic review of digital problem-based learning indicates that AI-based tools, including LLMs, can improve clinical learning by offering adaptive, personalized training environments that effectively engage healthcare professionals [48]. These findings indicate the potential of LLMs for enriching healthcare training and education.

Integrating LLMs into healthcare should align with Safety-II principles, which are focused on enhancing the ability of the system to succeed under varying conditions, rather than solely preventing errors. This approach recognizes that healthcare professionals constantly adjust their performance to match their work environment, an essential aspect of Safety-II [49]. While studies like Goh et al. demonstrate improved decision-making quality through physician-LLM collaboration [50], the actual value lies in how these tools support the adaptive capacity of healthcare professionals. This aligns with the Safety-II view of humans as crucial system flexibility and resilience resources. The goal should be to create collaborative human-AI partnerships that capitalize on both sources of strength and support the performance variability that Safety-II recognizes as necessary and beneficial in complex healthcare systems. Rather than following rigid protocols, AI tools should assist clinicians in adapting to evolving scenarios, with a focus on understanding and supporting Work-As-Done. This approach is consistent with the emphasis of Safety-II on creating conditions that enable people to perform successfully.

Furthermore, studies on AI-assisted documentation reveal that integrating LLMs can significantly reduce administrative burdens, such as the time nurses spend on routine documentation, freeing them to focus more on patient-centric care. Generative AI can automate tasks such as form-filling and EHR documentation, enhancing the efficiency of healthcare professionals. Generative AI systems can interpret unstructured data like medical notes and images, facilitating automated documentation, medical coding, billing, appointment scheduling, and patient inquiries, all through a user-friendly interface [51]. This allows healthcare providers to focus on direct patient care, ultimately improving patient experiences and operational efficiency. Similarly, AI-powered digital scribes using speech recognition have demonstrated promise in reducing clinician documentation efforts, though challenges remain in adapting to the complex dynamics of clinical settings [52]. The ultimate goal of AI integration should be to augment rather than replace human expertise, maintaining the critical human elements of patient care [53]. Therefore, the successful incorporation of LLMs into clinical workflows necessitates technological advancement coupled with a deep understanding of the socio-technical complexities of healthcare, in addition to ensuring user-friendly interfaces, and robust training.

1.9. Ethical considerations and integration into clinical workflows

A robust ethical framework is essential for responsible LLM deployment in healthcare. [54–56]. Central concerns include adhering to privacy and security regulations, such as HIPAA or GDPR, to protect patient data; mitigating inadvertent biases that may disproportionately affect certain populations by employing diverse training sets and federated learning; ensuring transparency and accountability so both clinicians and patients can understand and verify LLM outputs; and maintaining informed consent, particularly as AI-driven tools handle increasingly sensitive tasks. Additionally, strategies like Retrieval-Augmented Generation (RAG) are critical for reducing misinformation and hallucinations. Ultimately, proper governance, model validation, and ongoing audits help ensure these technologies enhance patient care without undermining trust or safety.

One primary concern involves privacy and data security, especially when integrating LLM-based agents into outbreak analytics and other healthcare applications. Robust data protection protocols are vital to safeguard sensitive health information, as these models are integrated into workflows that deal with sensitive patient information [57,58], to maintain strict privacy standards to avoid potential breaches and uphold patient trust [59]. Proper governance is another critical factor, ensuring that data ownership, patient consent, and the right to explanation for AI-assisted decisions are properly managed [60].

Bias and fairness in healthcare LLMs also raise substantial ethical issues. Research has revealed notable biases, such as those based on racial, ethnic, or gender differences, that can significantly affect health outcomes [61]. The implications of these biases are profound, especially in applications like pain management, where disparities in treatment recommendations could lead to inequitable care [57]. Studies suggest numerous strategies for mitigating these biases, including enhancing data diversity and adopting algorithms to recognize and correct unfair patterns [57]. Transparency and explainability are also essential features that must be incorporated into LLMs, particularly in the healthcare domain. The opaque nature of many LLMs complicates medical decision-making processes; clinicians and patients should benefit from clear, interpretable reasoning behind AI-generated suggestions [58].

Hallucinations—instances where an LLM generates incorrect or fabricated content—pose a serious risk in healthcare, where accuracy is paramount. [62] These inaccuracies may arise from insufficient domain-specific training data or the model's tendency to 'fill in gaps' when faced with unfamiliar queries. Several techniques have been proposed to reduce hallucinations and improve factual accuracy (see Table 2). Prompt engineering helps guide the model's outputs through carefully constructed queries and instructions. Retrieval-Augmented Generation (RAG) grounds the LLM's responses in a trusted external knowledge base, reducing reliance on "guesses" or incomplete training data. Finally, fine-tuning with domain-specific datasets can refine the model's knowledge, thereby enhancing factual reliability and reducing the incidence of spurious outputs. By combining these approaches, LLM-based systems can offer more dependable, evidence-based assistance to healthcare professionals and patients.

1.10. Future Directions

The evolution of medical LLMs is progressing, with multiple research areas emerging as critical for the responsible and effective deployment of these technologies in healthcare settings. [Table 3] summarizes Future Directions for Medical LLMs. One of the most crucial areas is interdisciplinary collaboration, which emphasizes building solid partnerships between AI researchers, healthcare professionals, ethicists, and policymakers. Such collaboration is necessary to ensure that LLMs are effective and ethically integrated into healthcare systems. Educational initiatives are being developed to bridge the gap between AI and medicine, aimed at ensuring that healthcare professionals have the necessary knowledge to use these advanced technologies effectively [63]. These programs are aimed at cultivating an AI-literate workforce that can leverage LLMs to improve patient care, while being mindful of the ethical concerns these technologies present, such as bias, privacy, and informed consent.

Another critical direction for medical LLMs is the development of new benchmarks that go beyond traditional accuracy metrics, aimed at incorporating safety, ethical, and clinical relevance measures. For example, new evaluation metrics must address how LLMs perform under diverse real-world conditions, including varying patient demographics, linguistic variations, and socio-economic contexts [58]. Clinically relevant benchmarks would help to ensure that these models not only deliver correct answers but also provide safe and culturally sensitive recommendations, particularly in high-stakes scenarios like diagnostics

Table 2
Summary of techniques to mitigate hallucination.

Technique	Definition	Benefits	Limitations
Prompt Engineering	The process of systematically designing and crafting prompts (queries, instructions, or input text) that guide an LLM's responses toward accurate and relevant outputs. This can include specifying context, constraints, or desired answer formats.	- Easy to Implement: Does not require modifying the model's architecture or datasets. - Flexible: Can be tailored to each new task or specialty.	- Trial-and-Error: Often requires iterative adjustments to prompts. - Limited Impact: Cannot fully correct deeper knowledge gaps in the model.
Retrieval-Augmented Generation (RAG)	A method that combines an LLM with an external, trusted knowledge source (e.g., medical databases, guidelines, or ontologies). The LLM retrieves" relevant facts and data to ground its responses, reducing reliance on purely predictive text generation.	- Improves Accuracy: Model relies on curated data rather than pure statistical guesswork. - Dynamic Updates: External data can be continually refreshed to ensure up-to-date knowledge.	- Dependency on Source Quality: Erroneous or outdated data in external databases can propagate into LLM outputs. - Complex Integration: Requires robust API or system design to retrieve relevant evidence.
Fine-Tuning	The process of training/re-training an LLM on high-quality, domain-specific datasets beyond its general-purpose pretraining. This improves the model with specialized knowledge and biases it toward more accurate outputs in that domain.	- Increased Factual Reliability: Model gains deeper, context-specific expertise. - Long-Term Enhancement: Once fine-tuned, the model consistently applies new knowledge.	- Resource-Intensive: Requires computational resources and carefully curated datasets. - Maintenance: Updated data may necessitate periodic re-fine-tuning.

or treatment planning. Additionally, creating standards for measuring ethical outcomes, such as fairness across demographic groups, would help mitigate risks related to health disparities and unequal treatment outcomes.

The field is also exploring multimodal LLMs (or large multimodal models, LMMs), which integrate diverse forms of data such as text, images, video, signal and structured patient information. These LLMs offer a holistic view of patient health by incorporating modalities like medical imaging alongside clinical notes, enabling more comprehensive decision-making [64]. These approaches are especially promising in radiology and nuclear medicine, where image-text domain conversion can streamline reporting and enhance diagnostic accuracy. [65] By interpreting images alongside clinical notes, LMMs may automate or assist in lesion detection, radiology reporting, and triaging, potentially improving workflow efficiency and reducing radiologist workload. Additionally, ongoing research focuses on refining the alignment between imaging features and textual outputs to mitigate errors and ensure clinical reliability. In the future, multimodal models will be like medical practitioners who need to use multiple data sources to make integrated disease judgments. This integration is particularly promising for applications in diagnostics, where combining visual information from imaging with text-based health records could improve the accuracy of disease detection and treatment planning. Cost-effective training methods for resource-intensive data types like medical imaging are also being developed, as computational costs constitute a significant barrier to implementing multimodal AI systems in healthcare settings.

The development of LLM-based medical agents for executing complex, multistep clinical tasks is increasingly becoming a research focus, as these agents transition from merely processing language to actively participating in clinical workflow. These agents could assist in real-time decision-making, such as in emergency department triage or creating personalized treatment plans during consultations, by leveraging their capability to interact autonomously with various databases, including EHRs and clinical guidelines [66]. Their evolving role could include multi-step analyses and collaborations with other agents, allowing them to support healthcare professionals in intricate tasks. Evaluating the effectiveness of these LLM agents requires frameworks like the "Artificial Intelligence Structured Clinical Examinations" (AI-SCE), which simulate real-world clinical environments to test their performance and adaptability. This assessment approach is focused on how well the agents handle complex scenarios, their interaction with users and tools, and their ability to perform in high-stakes, dynamic clinical settings. By incorporating continuous learning mechanisms, these LLM agents could stay updated with evolving medical guidelines, enhancing their relevance and reliability in environments that demand nuanced responsive medical interventions, such as during emerging health crises. To prevent overfitting—where a model memorizes a narrow question set rather than demonstrating broader, generalizable medical knowledge—AI-SCE design should include scenario-based questions, dynamic test banks, and real-world clinical vignettes. These strategies ensure that the examination evaluates the model's adaptability to diverse and evolving healthcare scenarios, rather than its capacity to memorize a limited set of questions.

Addressing underrepresented medical specialties, especially rare diseases, remains a challenge for the effective use of LLMs in healthcare. Rare diseases, affecting around 300 million people worldwide, are typically linked to small datasets due to their low prevalence and the complexity of data collection [67]. This scarcity complicates the training of LLMs, usually designed using more comprehensive medical datasets, resulting in insufficient representation of these specialized areas. There is an urgent need for improved data collection methods and inclusive training datasets encompassing rare and niche disease profiles to ensure equitable AI benefits across medical domains [68]. Recent advances, such as prompt learning and data augmentation techniques, have shown the potential to mitigate the challenges posed by small datasets. For instance, ChatGPT has demonstrated promising results in extracting rare disease phenotypes using zero-shot and few-shot learning with minimal labeled data. Additionally, adaptive research methodologies provide alternative ways to gather meaningful data from limited samples, which is critical for rare disease research [68]. Another promising avenue is federated learning, which allows multiple institutions to collaboratively train models without directly sharing patient data. By maintaining data locally and only transmitting model updates, federated learning preserves privacy while expanding training datasets. This approach is especially beneficial for rare diseases, where patient populations and data are scarce and often sparsely distributed another different data holders. Through privacy-preserving, multi-institutional collaboration, LLMs can gain exposure to more diverse examples, thus improving diagnostic accuracy and treatment recommendations for these specialized domains. By leveraging these approaches, researchers can enhance the adaptability of LLMs, enabling them to generalize effectively with limited training data. This is crucial for promoting equity in healthcare and helping to ensure that AI advances benefit all medical specialties, including those typically underrepresented.

The ethical development and responsible deployment of LLMs in healthcare is another crucial area. Accurately setting up and deploying explorations and constraints within existing knowledge requires continuous experimentation to effectively and correctly utilize LLMs. Ensuring that models mitigate biases, particularly in their training data, is essential to avoid perpetuating existing disparities in healthcare delivery [69]. Guidelines are also being established for transparency in LLM operations, making it possible for healthcare providers and patients

Table 3
Future directions for medical LLMs.

Research Area	Key Points	Challenges
Interdisciplinary Collaboration	Building partnerships between AI researchers, healthcare professionals, ethicists, and policymakers to help ensure ethical integration. Developing educational initiatives for AI literacy among clinicians [63].	Limited collaboration between diverse fields makes it difficult to establish shared ethical standards.
Development of New Benchmarks	Creating clinically relevant benchmarks that go beyond accuracy, including safety, ethics, and socio-cultural sensitivity. Establishing standards to help ensure fair outcomes [58].	Developing standardized benchmarks aimed at reflecting diverse real-world conditions and patient demographics.
Multimodal LLMs	Integrating multiple data types (e.g., text, images, signal, video) to provide a holistic view of patient health and improve diagnostics. Focus on cost-effective training for medical imaging [64].	High computational cost and resource-intensive data integration.
LLM-Based Medical Agents	Agents assist in real-time decision-making (e.g., triage, personalized treatments) and leverage databases for complex workflows. Use AI-SCE to evaluate agent performance in clinical environments [66].	Maintaining adaptability and accuracy in high-stakes scenarios, and managing data integration from multiple sources.
Addressing Underrepresented Specialties	Expanding LLM applications to rare diseases and niche specialties by enhancing data collection and using techniques like prompt learning and data augmentation [67,68].	The scarcity of data for rare diseases and underrepresented specialties limits training effectiveness.
Ethical Development & Transparency	Mitigating biases in training data, with the aim of ensuring that models do not perpetuate healthcare disparities. Developing guidelines for transparency and explainability to improve clinician trust [69].	Addressing inherent biases in training datasets, and creating explainable AI systems that are interpretable by non-experts.
Integration with Robotic Systems	Developing LLM-guided surgical robots and assistive devices to enhance precision, adaptability, and support for autonomous procedures. Integrate AR and haptic feedback for guidance [70].	Challenges in developing real-time response capabilities and ensuring patient-specific adaptability.

to understand how decisions are made and facilitating accountability.

Lastly, integrating LLMs with robotic systems represents an exciting frontier in healthcare technology. Integrating LLMs can significantly enhance the precision, adaptability, and decision-making capabilities of robotic surgical systems. For example, LLM-guided surgical robots have the potential to interpret a surgeon's spoken commands and assist by anticipating the next steps or adjusting their operations in real-time to meet patient-specific needs. LLMs can also support autonomous or semi-autonomous surgical procedures, where intelligent robotic systems could perform specific tasks with minimal human intervention, thereby reducing surgeon fatigue and potentially contributing to improvement in overall precision and patient outcomes [70]. Furthermore, natural language interfaces powered by LLMs can make medical equipment more accessible to healthcare providers, thereby reducing the learning curve for complex machinery and improving usability. These LLMs could also be integrated with augmented reality and haptic feedback systems to provide immersive guidance during surgery, which could enhance both the surgeon's situational awareness and procedure safety.

2. Conclusion

LLMs have demonstrated significant potential to revolutionize healthcare in their capacity to assist in the transformation of clinical decision-making, documentation, medical education, and research. These models have performed impressively on medical examinations, clinical reasoning tasks, and specialized diagnostics, often matching or surpassing human abilities, which indicates their capability to support healthcare professionals. LLMs also show promise with regard to enhancing patient communication by providing accurate, readable, and empathetic responses and streamlining clinical workflows through effective information extraction and medical record analysis. Despite these advances, integrating LLMs into healthcare is faced with challenges like hallucination, data limitations, ethical concerns, and regulatory complexities, which must be addressed through robust evaluation, improved adaptation methods, and alignment with medical practices. Ultimately, the successful deployment of LLMs will require interdisciplinary collaboration that emphasizes patient safety, ethical integrity, and human-centered implementation. By fostering partnerships between AI researchers, healthcare professionals, and policymakers, we can help to ensure these powerful AI tools enhance, rather than replace, human expertise and compassion in healthcare.

Acknowledgement

This research was supported by Chang Gung Memorial Hospital under grant numbers CORPG3L0461, CLRPG3H0016, and OMRPG3K0011, as well as by the National Science and Technology Council (NSTC), Taiwan, under grant number NSTC113-2640-B-182A-001. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] Qiu J, Lam K, Li G, Acharya A, Wong TY, Darzi A, et al. LLM-based agentic systems in medicine and healthcare. Nat Mach Intell 2024;6(12):1418–20.
[2] Chow JCL, Wong V, Li K. Generative pre-trained transformer-empowered healthcare conversations: current trends, challenges, and future directions in Large Language model-enabled medical chatbots. BioMedInformatics 2024;4(1):837–52.
[3] Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large Language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023;9:e48291.
[4] Demszky D, Yang D, Yeager DS, Bryan CJ, Clapper M, Chandhok S, et al. Using large language models in psychology. Nat Rev Psychol 2023;2(11):688–701.
[5] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv:2005.14165 2020.
[6] Siddique S, Chow JCL. Machine learning in healthcare communication. Encyclopedia 2021;1(1):220–39.
[7] Zhou H, Liu F, Gu B, Zou X, Huang J, Wu J, et al. A survey of Large Language Models in medicine: progress, application, and challenge. arXiv: 2311.05112 2023.
[8] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172–80.
[9] Chow JCL, Li K. Ethical considerations in human-centered AI: advancing oncology chatbots through Large Language Models. JMIR Bioinform Biotechnol 2024;5:e64406.
[10] Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling laws for neural Language Models;2020. https://doi.org/10.48550/arXiv.2001.08361.
[11] Naveed H, Ullah Khan A, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of Large Language Models. https://doi.org/10.48550/arXiv.2307.06435.
[12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. https://doi.org/10.48550/arXiv.1706.03762; 2017.

[13] Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, et al. Large Language models: a survey. arXiv:2402.06196 2024.
[14] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: scaling language modeling with pathways. arXiv:2204.02311 2022.
[15] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40.
[16] Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. https://doi.org/10.48550/arXiv.1904.05342.
[17] Yang X, Chen A, PourNejatian N, Shin HC, Smith EK, Parisien C, et al. GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records. arXiv:2203.03540 2022.
[18] Saab K, Tu T, Weng WH, Tanno R, Stutz D, Wulczyn E, et al. Capabilities of Gemini models in medicine. arXiv:2404.18416 2024.
[19] Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepano C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198.
[20] Tsoutsanis P, Tsoutsanis A. Evaluation of Large language model performance on the multi-specialty recruitment assessment (MSRA) exam. Comput Biol Med 2024;168:107794.
[21] Nakajima N, Fujimori T, Furuya M, Kanie Y, Imai H, Kita K, et al. A comparison between GPT-3.5, GPT-4, and GPT-4V: can the Large Language model (ChatGPT) pass the Japanese board of orthopaedic surgery examination? Cureus 2024;16(3):e56402.
[22] Park L, Ehlert B, Susla L, Lum ZC, Lee PK. Performance of large language model artificial intelligence on dermatology board exam questions. Clin Exp Dermatol 2024;49(7):733–4.
[23] Williams CYK, Zack T, Miao BY, Sushil M, Wang M, Kornblith AE, et al. Use of a Large Language model to assess clinical acuity of adults in the emergency department. JAMA Netw Open 2024;7(5):e248895.
[24] Young CC, Enichen E, Rivera C, Auger CA, Grant N, Rao A, et al. Diagnostic accuracy of a custom Large Language model on rare pediatric disease case reports. Am J Med Genet 2025;197(2):e63878.
[25] Ye C, Zweck E, Ma Z, Smith J, Katz S. Doctor versus artificial intelligence: patient and physician evaluation of Large Language model responses to rheumatology patient questions in a cross-sectional study. Arthritis Rheumatol 2024;76(3):479–84.
[26] Shin E, Hartman M, Ramanathan M. Performance of the ChatGPT large language model for decision support in community pharmacy. Br J Clin Pharmacol 2024;90(12):3320-33.
[27] Zhou J, He X, Sun L, Xu J, Chen X, Chu Y, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. Nat Commun 2024;15(1):5649.
[28] Zhang L, Liu M, Wang L, Zhang Y, Xu X, Pan Z, et al. Constructing a Large Language model to generate impressions from findings in radiology reports. Radiology 2024;312(3):e240885.
[29] Zheng C, Ye H, Guo J, Yang J, Fei P, Yuan Y, et al. Development and evaluation of a large language model of ophthalmology in Chinese. Br J Ophthalmol 2024;108(10):1390–7.
[30] Gao Y, Li R, Emma C, Caskey J, Brian WP, Matthew C, Miller T, et al. Leveraging A medical knowledge graph into Large Language Models for diagnosis prediction. arXiv:2308.14321 2023.
[31] Zhou S, Luo X, Chen C, Jiang H, Yang C, Ran G, et al. The performance of large language model powered chatbots compared to oncology physicians on colorectal cancer queries. Int J Surg 2024;110(10):6509–17.
[32] Halawani A, Mitchell A, Saffarzadeh M, Wong V, Chew BH, Forbes CM. Accuracy and readability of kidney stone patient information materials generated by a Large Language model compared to official urologic organizations. Urology 2024;186:107–13.
[33] Kozaily E, Geagea M, Akdogan ER, Atkins J, Elshazly MB, Guglin M, et al. Accuracy and consistency of online large language model-based artificial intelligence chat platforms in answering patients' questions about heart failure. Int J Cardiol 2024;408:132115.
[34] Small WR, Wiesenfeld B, Brandfield-Harvey B, Jonassen Z, Mandal S, Stevens ER, et al. Large Language model-based responses to patients' in-basket messages. JAMA Netw Open 2024;7(7):e2422399.
[35] He W, Zhang W, Jin Y, Zhou Q, Zhang H, Xia Q. Physician versus Large Language model chatbot responses to web-based questions from autistic patients in Chinese: cross-sectional comparative analysis. J Med Internet Res 2024;26:e54706.
[36] Horváth Á, Molnár P. A review of patient safety communication in multicultural and multilingual healthcare settings with special attention to the U.S. and Canada. Dev Health Sci 2021;4(3):49–57.
[37] Qiu P, Wu C, Zhang X, Lin W, Wang H, Zhang Y, et al. Towards building multilingual language model for medicine. Nat Commun 2024;15(1):8384.
[38] Sulejmani P, Negris O, Aoki V, Chu CY, Eichenfield L, Misery L, et al. A large language model artificial intelligence for patient queries in atopic dermatitis. J Eur Acad Dermatol Venereol 2024;38(6):e531–5.
[39] Omar M, Soffer S, Charney AW, Landi I, Nadkarni GN, Klang E. Applications of large language models in psychiatry: a systematic review. Front Psychiatry 2024;15:1422807.
[40] Sushil M, Zack T, Mandair D, Zheng Z, Wali A, Yu YN, et al. A comparative study of large language model-based zero-shot inference and task-specific supervised classification of breast cancer pathology reports. J Am Med Inform Assoc 2024;31(10):2315–27.
[41] Yasaka K, Kanzawa J, Kanemaru N, Koshino S, Abe O. Fine-tuned Large odel for extracting patients on pretreatment for lung cancer from a picture archiving and communication system based on radiological reports. J Imaging Inform Med 2025;38(1):327-34.
[42] Laukamp KR, Terzis RA, Werner JM, Galldiks N, Lennartz S, Maintz D, et al. Monitoring patients with glioblastoma by using a Large Language model: accurate summarization of radiology reports with GPT-4. Radiology 2024;312(1):e232640.
[43] Heilmeyer F, Böhringer D, Reinhard T, Arens S, Lyssenko L, Haverkamp C. Viability of open Large Language Models for clinical documentation in German health care: real-world model evaluation study. JMIR Med Inform 2024;12:e59617.
[44] Ontika N, Saßmannshausen S, Syed H, de Carvalho A. Exploring human-centered AI in healthcare: a workshop report. In: Volkmar Pipek, Markus Rohde. International Institute for Socio-Informatics (IISI; 2022. 19(2).
[45] Waefler T, Schmid U. Explainability is not enough: requirements for human-AI-partnership in complex socio-technical systems;. 2021. https://doi.org/10.34190/EAIR.20.007.
[46] Bienefeld N, Keller E, Grote G. Human-AI teaming in intensive care: a socio-technical systems view and international delphi study among data scientists (preprint). 2023. https://doi.org/10.2196/preprints.50130.
[47] Waisberg E, Ong J, Masalkhi M, Lee AG. Large language model (LLM)-driven chatbots for neuro-ophthalmic medical education. Eye (Lond) 2024;38(4):639–41.
[48] Tudor Car L, Kyaw BM, Dunleavy G, Smart NA, Semwal M, Rotgans JI, et al. Digital problem-based learning in health professions: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019;21(2):e12945.
[49] Hollnagel E, Wears R, Braithwaite J. From safety-I to safety-II. A White Paper; 2015. https://doi.org/10.13140/RG.2.1.4051.5282.
[50] Goh E, Gallo R, Strong E, Weng Y, Kerman H, Freed J, et al. Large Language model influence on management reasoning: a randomized controlled trial. medRxiv 2024. https://doi.org/10.1101/2024.08.05.24311485.
[51] Falak PDL. A survey on healthcare virtual assistant using generative AI. Int J Res Appl Sci Eng Technol 2023;11(11):130–4.
[52] Quiroz JC, Laranjo L, Kocaballi AB, Berkovsky S, Rezazadegan D, Coiera E. Challenges of developing a digital scribe to reduce clinical documentation burden. NPJ Digit Med 2019;2:114.
[53] Bossen C, Pine KH. Batman and robin in healthcare knowledge work: human-AI collaboration by clinical documentation integrity specialists. ACM Trans Comput Hum Interact 2023;30(2):1–29.
[54] Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). NPJ Digit Med 2024;7(1):183.
[55] Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review. Diagn Pathol 2024;19(1):43.
[56] Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. Front Artif Intell 2023;6:1166014.
[57] Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023;25:e48009.
[58] Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine 2023;90:104512.
[59] Choudhury A, Chaudhry Z. Large Language models and user trust: consequence of self-referential learning loop and the deskilling of healthcare professionals. J Med Internet Res. 2024;26:e56764.
[60] Okonji OR, Yunusov K, Gordon B. Applications of Generative AI in Healthcare: algorithmic, ethical, legal and societal considerations. arXiv:2406.10632 2024.
[61] Jiao J, Afroogh S, Xu Y, Phillips C. Navigating LLM ethics: advancements, challenges, and future directions. arXiv:2406.18841 2024.
[62] Bhayana R. Chatbots and Large Language Models in radiology: a practical primer for clinical and research applications. Radiology 2024;310(1):e232756.
[63] Sblendorio E, Dentamaro V, Lo Cascio A, Germini F, Piredda M, Cicolini G. Integrating human expertise & automated methods for a dynamic and multi-parametric evaluation of large language models' feasibility in clinical decision-making. Int J Med Inf 2024;188:105501.
[64] Zhang C, Chen J, Li J, Peng Y, Mao Z. Large language models for human–robot interaction: a review. Biomim Intell Robot 2023;3(4):100131.
[65] Hirata K, Matsui Y, Yamada A, Fujioka T, Yanagawa M, Nakaura T, et al. Generative AI and large language models in nuclear medicine: current status and future prospects. Ann Nucl Med 2024;38(11):853–64.
[66] Mehandru N, Miao BY, Almaraz ER, Sushil M, Butte AJ, Alaa A. Evaluating large language models as agents in the clinic. NPJ Digit Med 2024;7(1):84.
[67] Shyr C, Hu Y, Bastarache L, Cheng A, Hamid R, Harris P, Xu H. Identifying and Extracting Rare Diseases and Their Phenotypes with Large Language Models. J Healthc Inform Res 2024;8(2):438–61.
[68] Mitani AA, Haneuse S. Small data challenges of studying rare diseases. JAMA Netw Open 2020;3(3):e201965.
[69] Al Nazi Z, Peng W. Large language models in healthcare and medical domain: a review. arXiv:2401.06775 2012.
[70] Ray PP. Large language models in laparoscopic surgery: A transformative opportunity. Laparoscopic, endoscopic and robotic. surgery 2024;7(4):174–80.
[71] Alfertshofer M, Knoedler S, Hoch CC, Cotofana S, Panayi AC, Kauke-Navarro M, et al. Analyzing question characteristics influencing ChatGPT's performance in 3000 USMLE®-Style questions. Med Sci Educ 2024;35(1):257–67.