International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

Comprehensive Literature Review: Large Language Models in Medical Decision Support Systems

G Anitha,Professor, Dept of CSE(AI&ML), CMR Engineering College, Telangana India, drganitha@cmrec.ac.in

Nakka Venkatesh, Assistant Professor, Dept of CSE(AI&ML), CMR Engineering College, Telangana India, venkatesh.n@cmrec.ac.in

Abstract- **Large Language Models (LLMs) have become a game-changing artificial intelligence (AI) technology that can revolutionize clinical decision support systems (CDSS) by leveraging its ability to process and understand natural language, to reason, and to interpret multimodal data. This review offers a detailed analysis of how LLMs are currently applied to medical decision-making tools including diagnosis, treatment planning, interpretation of medical imaging, clinical documentation, medication safety and personalized healthcare suggestions. This study aims to critically discuss the recent developments regarding Retrieval-Augmented Generation (RAG), Chain-of-Thought (CoT) reasoning, knowledge graph incorporation, multimodal fusion, and multi-agent collaborative AI systems, which are all aimed at boosting the clinical accuracy, minimizing hallucinations, and increasing explainability of AI systems. Moreover, the review explores the key limitations of LLMs in healthcare, such as potential biases, hallucinations, interpretability, regulatory compliance, ethical issues, privacy protection, and robustness in face of dataset shift. A comparative study of current medical LLM models reveals that, although these systems offer substantial advances in diagnostic reasoning and streamlining workflows, the role of the clinician remains indispensable, highlighting the need for a human-centric approach to AI in healthcare. Finally, future research directions are discussed to facilitate next-generation intelligent healthcare infrastructures for reliable and clinically deployable applications: uncertainty-aware reasoning, domain-specialized medical LLMs, trustworthy explainable AI, federated medical LLMs, and collaborative multi-agent healthcare systems**.

Keywords : *Large Language Models, Clinical Decision Support Systems, Retrieval-Augmented Generation, Explainable Artificial Intelligence, Multimodal Medical AI, Medical Natural Language Processing, Healthcare AI Systems, AI-Assisted Clinical Reasoning*

I. INTRODUCTION AND FOUNDATIONAL CONCEPTS

The integration of Large Language Models (LLMs) into healthcare is a revolutionary technological development that profoundly impacts the way medical practitioners access information, make clinical decisions and provide patient care[1]-[3]. These systems, built on the basis of powerful neural networks and trained on a large volume of textual data are able to display extraordinary abilities in recognizing clinical natural language and generating it, making them excellent candidates as a component in clinical decision support systems (CDSS) [4],[5]. The distinction lies in the ability of LLMs to learn from extensive, unstructured clinical information such as electronic health records, Unlike rule-based systems with fixed logic, it provides a more nuanced and context-aware view which is useful for medical decision-making, radiology reports and patient narratives. This shift from conventional rule-based expert systems to the present-day LLMs represents a paradigm change in AI-driven healthcare. A large number of rules were needed to be formulated in such a traditional CDSS, and an iterative process of collaboration between clinicians and developers was time consuming and costly [2]. Modern LLMs can address these issues by directly acquiring intricate patterns from the training data, demonstrating outstanding performance in medical knowledge and clinical reasoning tests [8],[3]. They are not confined to text analysis, but also medical images, physiological signals and structured data from EHRs, leading to full fledged clinical decision support architectures [9].

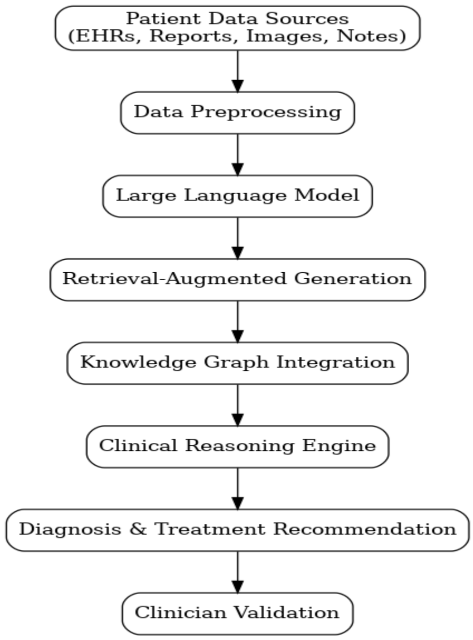

Fig 1: Architecture of LLM-Based Clinical Decision Support System

The LLM-based clinical decision support system (CDSS) for healthcare decision-making with electronic health records (EHR) and medical imaging, retrieval-augmented generation (RAG), and clinical reasoning modules is illustrated in Figure 1.

II. CLINICAL APPLICATIONS AND DIAGNOSTIC PERFORMANCE

A. Diagnostic Accuracy and Triage Support

Modern LLMs have demonstrated remarkable diagnostic abilities in the various fields of medicine with some models now performing at, or near, the level of a clinician on standardized tests [10],[8]. One of the first assessments of eight modern LLMs across a range of triage levels using clinical vignettes found that structured prompting substantially increased the accuracy of the LLMs in the diagnostics (93.75%) and triage (86.20%) tasks [8]. These systems are good at synthesising patient information, analysing medical literature and formulating differential diagnoses, which is important when recognising rare diseases, where pattern recognition across subtle combinations of symptoms is crucial [3].

LLMs have significantly increased the capabilities in medical imaging-based diagnosis. Multimodal LLMs combine vision transformers with natural language processing, which has shown great success in detecting abnormalities from radiographs, chest computed tomography scans, electrocardiograms and pathological images[11],[3]. LLMs have been shown to perform at or near human levels on board-style exams and subspecialty diagnostic reasoning in ophthalmology with a high degree of accuracy in diagnosis of common diseases and are fairly consistent in performance across user domains [12].

B. Treatment Planning and Medication Management

By leveraging patient-specific information and clinical guidelines, LLM can significantly aid in evidence-based treatment recommendations. In a multi-specialty evaluation, the LLM-based CDS tools were shown to substantially enhance medication safety in 16 clinical specialties, including a 1.5-fold improvement in accuracy relative to pharmacist review in identifying serious harm[13]. These systems can help the clinician in providing personalized treatment recommendations, while keeping track of drug-drug interactions, contraindications and adherence to guidelines [13].

LLMs can be used to provide assistance in medical treatment planning for specific areas of specialization. In oncology, LLMs are used in complex clinical decision making throughout the precision medicine continuum including cancer screening and diagnosis, staging, treatment recommendation and clinical documentation [14]. When combined with domain-specific knowledge and clinical practice guidelines, LLMs can offer insights for personalized medical treatment, patient risk evaluation, and optimization of therapeutic strategies in nephrology and cardiology [15],[16].

C. Clinical Documentation and Administrative Support

Besides supporting diagnostics and treatments, LLMs significantly improve healthcare workflows by enabling automated documentation and clinical summarisation. These systems can be used to automatically produce detailed clinical notes, discharge summaries, and progress reports based on the interactions between patients and doctors or structured patient data to ease the administrative burden and free up doctors' time for more direct patient care [17],[9]. These natural language processing features allow for the quick and efficient creation of complex clinical narratives, the ability to easily extract important information from long documents, and pattern matching across cohorts of patients, which is very useful in busy clinical environments where there is a lot of documentation to be done.

III. TECHNICAL APPROACHES AND ENHANCEMENT STRATEGIES

A. Retrieval-Augmented Generation (RAG)

To tackle some of the fundamental issues of LLMs, including stale information and hallucination, a groundbreaking technical solution, called Retrieval-Augmented Generation (RAG) has been proposed, which anchors LLM outputs in trusted sources of external knowledge [18],[19]. To make the most of LLMs to retrieve and interpret relevant medical text, approaches such as Retrieval-Augmented Generation (RAG) have been developed, which integrate document retrieval systems with generative models. To enable LLMs to access and draw insights from relevant medical documents, frameworks such as RAG have been introduced, which integrate document retrieval systems with generative models.

.TABLE 1: COMPARATIVE ANALYSIS OF LLM TECHNIQUES IN HEALTHCARE

Technique	Purpose	Advantages	Limitations
RAG	Knowledge Retrieval	Reduces hallucinations	Retrieval latency
CoT Prompting	Stepwise reasoning	Better interpretability	Higher computation
Knowledge Graphs	Domain grounding	Improved explainability	Complex integration
Multimodal LLMs	Multi-source learning	Better diagnosis	Large resource usage

Table 1 compares major LLM enhancement strategies in healthcare applications, highlighting their advantages, clinical relevance, and associated limitations.

RAG's effectiveness in medical use cases is proven. When used to interpret hepatological clinical guidelines, GPT-4 Turbo's performance was compared to a coherent text format of the guidelines plus few-shot learning and rose from 43% to 99% accuracy [19]. Hybrid RAG models that integrate dense and sparse vector representations scored 0.89 F1 and 0.89 accuracy on test sets for pulmonary disease

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

prediction, significantly surpassing other deep learning models such as BERT [20].

In the context of orthopedic patient education, RAG-based chatbots scored high in accuracy (4.55/5), helpfulness (4.61/5) and ease of use (4.90/5) according to the specialists, students and non-medical users, and automated evaluation showed excellent performance in answer relevancy, contextual accuracy and faithfulness [21].

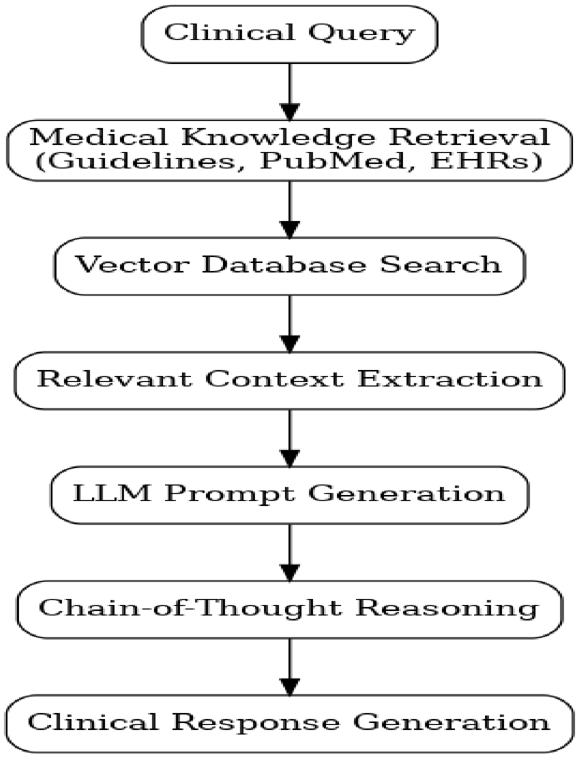

Fig 2: Workflow of Retrieval-Augmented Medical LLM Framework

Figure 2 illustrates the retrieval-augmented generation workflow where clinical queries are enriched using medical knowledge retrieval, vector database search, contextual reasoning, and large language model-based response generation.

Algorithm : Retrieval-Augmented Medical Decision Support

Input:
Clinical query Q
Medical knowledge database D
Output:
Clinical recommendation R

1: Receive clinical query
2: Retrieve relevant medical documents
3: Generate contextual embeddings
4: Construct augmented prompt
5: Apply LLM reasoning
6: Generate clinical response
7: Validate output using guidelines
8: Return recommendation

The proposed Algorithm describes the retrieval-augmented medical decision support process for generating evidence-grounded clinical recommendations using large language models and external medical knowledge repositories.

B. Prompt Engineering and Chain-of-Thought Reasoning

Structured prompting techniques, especially chain-of-thought (CoT) methods that break down clinical reasoning into logical steps, have significantly improved the accuracy of LLMs in diagnostics and the clarity of reasoning presented [20],[8]. Expert guided prompting given in specific clinical domains and certification levels further enhanced the performance—in the case of questions from emergency medical services, Expert-CoT achieved 2.05% accuracy improvement over vanilla CoT and Expert-CoT + Expert RAG showed 4.59% accuracy improvement over vanilla RAG baseline [22]. The use of summarisation methods using the TF-IDF and k-means clustering improves the LLMs understanding of radiologic findings specific to the disease [20]. These prompt engineering techniques are especially useful for noisy or long clinical narratives with a complex structure, in which early retrieval can help to ground the next steps of reasoning in domain facts, yet RAG with CoT is more effective for high-quality notes, where sequential reasoning can be of maximal benefit [23].

C. Knowledge Graph Integration and Multimodal Fusion

By combining knowledge graphs with LLMs synergistically, powerful systems are developed, not only boosted by domain knowledge but also better explainable. To overcome the hallucination and lack of domain grounding, knowledge graphs can be leveraged to facilitate structured medical relationships and definitions of entities [24],[25]. The use of this approach in a mental health CDSS resulted in an accuracy improvement of up to 12.0% and an F1 score improvement of up to 8.6% compared to standalone LLMs, and even more when combined with medical domain

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

knowledge [25]. Multimodal approaches that combine imaging data with structured clinical variables enable uncertainty-aware frameworks for robust clinical deployment [11],[9]. Regarding diabetes treatment, integrated image-language systems that integrate LLM modules with image-based deep learning gave personalized treatment advice, and the LLM exhibited similar performance to endocrinology residents in English, but superior performance in Chinese translations [11].

Algorithm: Multimodal LLM-Based Diagnostic Framework

Input:
Medical images, EHRs, clinical notes
Output:
Diagnostic prediction

1: Collect multimodal patient data
2: Extract image features
3: Process clinical text
4: Fuse multimodal representations
5: Apply LLM reasoning
6: Generate diagnosis and explanation
7: Produce treatment recommendation
8: Return final decision

The proposed Algorithm presents a multimodal large language model framework integrating medical imaging, clinical text, and electronic health records for intelligent diagnostic prediction and treatment recommendation.

IV. CHALLENGES AND SAFETY CONSIDERATIONS

A. Hallucinations, Bias, and Reliability Issues

LLMs have achieved impressive capabilities, but there are also significant limitations that hinder clinical deployment. Hallucinations, the production of believable but false information, are a basic problem in medical applications where accuracy is essential [26]-[28]. All 10 mainstream LLMs tested showed systematic, implicit gender, race, socioeconomic status and health biases, especially in the categories of Race (Mean=0.61) and Socioeconomic Status (Mean=0.56) [27]. Importantly, stronger implicit associations strongly predicted the discriminatory choices in downstream medical decision tasks ($p<0.001$) that can directly influence health equity and equitable patient care [27].

TABLE 2: CHALLENGES AND SOLUTIONS IN MEDICAL LLM DEPLOYMENT

Challenge	Impact	Possible Solution
Hallucination	Incorrect diagnosis	RAG + Verification
Bias	Health inequality	Fairness-aware training
Lack of Explainability	Low clinician trust	XAI integration
Privacy Risks	Regulatory issues	Federated Learning
Dataset Shift	Poor generalization	Uncertainty modeling

Table 2 outlines key technical, ethical, and operational considerations for implementing large language models in healthcare contexts, along with potential solutions.

The ramifications go beyond factual correctness to consistency and reliability. A study on metacognitive skills found that LLM was highly accurate on multiple choice medical questions, however, it was found to be very low on the ability to admit that it does not know the answer, and so it gave the “confident” answer even if no correct answer options were given [29]. The identification of strengths and weaknesses by the perceptions underscores potential challenges in clinical practice, where trust in AI-generated suggestions could lead to incorrect diagnoses or treatment mistakes.

B. *Interpretability and Explainability*

Another key issue with LLM integration in clinical workflows is that of interpretability. Traditional rule-based systems specify their decision making explicitly, however, for LLM decisions, it is not possible to explain the reasoning, which directly harms clinician trust and acceptance [30],[31]. Partial explanations are provided by methods such as attention mechanisms, SHAP, LIME and Grad-CAM visualization, but they do not fully represent the reasoning mechanisms behind the output of the LLM [32],[33].

A lack of good interpretability poses special problems in high-stakes medical settings. Clinicians should be able to interpret AI suggestions in a clear manner and have them resonate with their existing medical expertise, allowing them to either accept or adapt AI suggestions for the case at hand, while retaining full responsibility for clinical decisions [38],[31]. When integrating XAI techniques into clinical workflows, it is important to consider the users' needs and ensure that the explanations are actionable and easy to understand, without adding to the complexity of the workflow [31].

C. Data Quality and Generalization

The reliability of LLMs in healthcare applications is severely limited by data quality issues. The unstructured narratives in the EHR have variable annotation standards, missing values, and inconsistencies in documentation, all of which directly impact model performance [34],[35]. However, model generalization across different clinical populations and healthcare settings is still an issue, as models tend to lose accuracy when they are presented with patient populations, disease presentations, or clinical contexts that are different from those in the training data [30].

Robustness under dataset shift, which is when patients seen in practice may differ from those in the training set or when patients are viewed in different clinical settings, is an

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

important part of clinical utility [36]. One way to improve robustness is to pair uncertainty estimation and multimodal learning, so that models can raise alarms for high-risk predictions that require expert review, but not give false assurances to clinicians for substandard predictions [37,36].

V. ETHICAL, REGULATORY, AND IMPLEMENTATION CONSIDERATIONS

A. Ethics, Privacy, and Trustworthiness

When considering the ethical implications of using LLMs in healthcare, numerous challenges arise, including data privacy, informed consent, algorithmic fairness, and accountability frameworks [38],[5]. The sensitivity and regulatory standards of healthcare, especially in the context of GDPR, HIPAA, and new regulations such as the EU AI Act [34],[39], highlight the need for patient privacy. Technical solutions, such as homomorphic encryption, differential privacy and federated learning, offer possibilities to ensure privacy, but are challenging to achieve in a healthcare context [34],[38]. Accountability mechanisms are not well developed in clinical AI applications. CDSS recommendations based on LLM, which resulted in harmful clinical outcomes, have unclear lines of responsibility, and it is uncertain who is at fault—the clinician, the healthcare institution, or the technologic developer [39],[35]. For the ethical implementation, there needs to be clear governance structures, human responsibility is defined, and there is space for suitable AI-human collaboration[39].

B. Regulatory Frameworks and Compliance

New regulations, such as the European Union's AI Act and regulatory guidance in the United States, have added requirements for transparency and safe validation and post-market surveillance of medical AI systems [39,40]. Such organizations are subject to double regulatory requirements because LLM systems for clinical decision support are often regulated as clinical decision support software or medical devices, and therefore have multiple validation requirements [38]. Innovation and patient protection need to be balanced in the regulatory compliance field. The EU AI Act provides for risk-based strategies and more stringent measures for higher-risk medical uses, including transparency, human oversight, monitoring for bias, and strong governance mechanisms [39,[40]. Organizations need holistic quality management systems that cover data governance, algorithm transparency, constant monitoring, and a way to recalibrate the algorithms when their performance begins to wane over time [40].

C. Human-Centered Design and Clinical Workflow Integration

Human centered design principles that focus on the needs of the clinicians and systems that support and augment existing workflow are essential to successful LLM deployment [7],[41]. LLM-augmented AI CDSS trials have shown that the provision of a response by the LLM that includes citations can help build trust over the AI CDSS without citations, while human-computer interaction can vary significantly depending on clinical expertise [41].

Implementing LLMs optimally involves using them as team members, not decision-making agents [41]. Effective deployment needs iterative design cycles with clinician feedback, an understanding of alert fatigue and automation bias, as well as ongoing education of clinicians about the capabilities and limitations of the AI systems [4],[42].

VI. SPECIALTY-SPECIFIC APPLICATIONS AND FUTURE DIRECTIONS

A. Domain-Specific LLM Development

By fine-tuning a general LLM on medical information and guidelines, specialized models have shown promise in outperforming general models. The ophthalmology-specific LLM, EyeGPT, achieved much higher performance in supplying informed advice and minimizing hallucination by fine-tuning based on ophthalmology-related datasets and retrieval augmentation from ophthalmology textbooks [43].

In the same way, LiVersa, an LLM specialized in liver disease, showed greater accuracy than ChatGPT-4 and other general LLMs in an evaluation by hepatologists on hepatology-related topics, but also presented some limitations that need further development for clinical use [44]. Multinational medical bodies are turning to LLM assistants powered by RAG, optimized for particular clinical areas and organizational settings. The Hematobot platform built by the Spanish Society of Hematology (SEH) used RAG pipelines and embraced the usage of GPT-4.1, with a set of rules ensuring adherence to guidelines, dosing precautions, and proper citations in several myeloma management and supportive care documents [45]. These specialized systems are shown to have significantly better performance on clinician-validated question sets, but still need ongoing improvement and evaluation when clinical evidence changes.

Multi-Agent Systems and Collaborative AI

Multiple specialized LLM agents showed improved performance on complex clinical reasoning tasks in advanced frameworks. These complex problems of diagnostic and treatment planning can be broken down into smaller, specialized tasks, such as multiple agents making decisions about patient admission (triage agents), agents making primary decisions (diagnosis agents), and agents giving personalized recommendations (treatment agents) [6]. This MAP (Multi-Agent Inpatient Pathways) framework resulted in an increase of 25.10% in diagnosis accuracy when compared to baseline

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

LLMs, outperforming three board-certified clinicians by 10% - 12% [6].

Autonomous retrieval and re-ranking systems with agentic memory retrieval capabilities that use specialized tools for retrieving, re-ranking, and providing grounding of evidence evidence are significantly better than standalone LLMs. On various medical QA metrics, unified agentic systems reached 82.98% accuracy in USMLE Step 1 and 86.24% accuracy in USMLE Step 2, outperforming GPT-4's performance and closely competing with the best proprietary medical LLM systems [46].

B. Emerging Research Directions

Several key areas where future developments of LLM should be focussed. It is important to recognize that enhanced uncertainty quantification – distinguishing between epistemic uncertainty (knowledge gaps) and aleatoric uncertainty (irreducible noise) – can help a clinician identify those cases that need to be brought to the attention of experts [25],[36]. Raising metacognitive skills—on LLM's ability to identify and communicate knowledge gaps—will significantly contribute to clinical reliability [29].

Another priority is research on the robustness of datasets and real world robustness. The variability in clinical settings is enormous and patient characteristics as well as disease manifestation and clinical situation varies constantly, not fully captured by training data[47],[37]. Efficient mechanisms for ongoing learning and performance monitoring and methods that would allow safe deployment in resource-limited healthcare environments that do not have high-end infrastructure would lead to global deployment.

C. Research gaps and Open Challenges

While significant advancements have been achieved in medical large language models, several important challenges have yet to be addressed, such as the potential for hallucination, the need for interpretability, the issue of bias in datasets, the burden of computation, privacy concerns, and sub-optimal robustness in the presence of real-world dataset shift dynamics. Moreover, current systems lack the ability for uncertainty-aware reasoning, multimodal scalability, explainability in high-stakes applications, and regulatory approval for clinical applications.

VII. CONCLUSION

Overall, LLMs offer a promising technology with significant potential to improve medical decision support, the accuracy of diagnoses, treatment planning, and clinical workflows. Recent significant developments in the incorporation of retrieval-augmented generation (RAG), prompt engineering, knowledge graph structures, and multimodal learning have significantly overcome basic limitations and enhanced accuracy, explainability, and minimization of hallucinations. But there are still major issues in terms of bias, interpretability, data quality, regulatory compliance and clinical integration. To realize the full potential of LLMs and their applications in healthcare, policy makers, clinicians, AI researchers, and ethicists must continue to work together in an interdisciplinary manner, carry out rigorous validation in clinical environments, and adopt robust governance structures that balance innovation with patient protection, and design AI systems for clinical use with a human-centric approach that complements, rather than replaces, medical expertise. LLMs hold the potential to greatly transform healthcare delivery by supporting more precise, efficient, and equitable clinical decision-making processes, all while preserving the irreplaceable nature of human clinical judgment, and, with suitable protection measures and ongoing improvements, can be integrated into the clinical workflow.

REFERENCES

[1] U. Sarker, A. Biswas, Y. K. Rauniyar, D. K. Jha, and A. D. Shreshtha, "Explainable large language models for clinical decision support: Multi-modal deep learning protocol with neuroscience and electronic health records," *International Journal of Innovative Science and Research Technology*, 2026, doi: 10.38124/ijisrt/26feb1223.

[2] S. Sivasothy, A. Bingham, I. Logothetis, S. Barnett, M. Abdelrazek, C. Luckhoff, J. Mathew, R. Vasa, and K. Mouzakis, "Large language models and conditional rules in clinical decision support systems," *Health Information Science and Systems*, 2026, doi: 10.1007/s13755-026-00428-z.

[3] X. Yang *et al.*, "Application of large language models in disease diagnosis and treatment," *Chinese Medical Journal*, 2024, doi: 10.1097/CM9.0000000000003456.

[4] S. Maity and M. Saikia, "Large language models in healthcare and medical applications: A review," *Bioengineering*, 2025, doi: 10.3390/bioengineering12060631.

[5] A. Shuja, "Artificial intelligence in clinical decision making: Opportunities and ethical challenges," *Developmental Medico-Life-Sciences*, 2026, doi: 10.69750/dmls.03.02.0192.

[6] Z. Chen, Z. Peng, X. Liang, C. Wang, P. Liang, L. Zeng, M. Ju, and Y. Yuan, "MAP: Evaluation and multi-agent enhancement of large language models for inpatient pathways," *arXiv*, 2025, doi: 10.48550/arXiv.2503.13205.

[7] M. B. Kola, "Integration of large language models in clinical decision support: A framework for human-AI collaboration in healthcare," *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, 2024, doi: 10.32628/cseit2410612432.

[8] L. Xu, W. Zhao, and X. Huang, "Diagnosis and triage performance of contemporary large language models on short clinical vignettes," *Journal of Medical Systems*, 2025, doi: 10.1007/s10916-025-02284-y.

[9] Q. Niu *et al.*, "From text to multimodality: Exploring the evolution and impact of large language models in medical practice," *arXiv*, 2024, doi: 10.48550/arXiv.2410.01812.

[10] T. Tu *et al.*, "Towards conversational diagnostic artificial intelligence," *Nature*, 2025, doi: 10.1038/s41586-025-08866-7.

[11] J. Li *et al.*, "Integrated image-based deep learning and language models for primary diabetes care," *Nature Medicine*, 2024, doi: 10.1038/s41591-024-03139-8.

[12] L. Cohen, A. R. Gupta, P. Patel, G. S. Gill, H. K. Bains, and S. K. Gupta, "The role of large language models in ophthalmology: A review of current applications, performance, and future directions," *Cureus*, 2025, doi: 10.7759/cureus.97374.

[13] J. Ong *et al.*, "Large language model as clinical decision support system augments medication safety in 16 clinical specialties," *Cell Reports Medicine*, 2025, doi: 10.1016/j.xcrm.2025.102323.

[14] X. Guo *et al.*, "Precision oncology: From large language models to multi-agent systems," *Frontiers in Oncology*, 2026, doi: 10.3389/fonc.2026.1828507.

[15] K. Yang *et al.*, "Natural language processing for cardiology: A narrative review," *arXiv*, 2025, doi: 10.48550/arXiv.2510.16708.

[16] C. Zoccali, L. Floyd, O. Cseprekál, M. Eisenga, Ş. Mirioğlu, F. Caravaca-Fontán, and F. Mallamaci, "AI-driven nephrology: The role of large language models in kidney care," *American Journal of Nephrology*, 2025, doi: 10.1159/000548208.

[17] M. Cascella, F. Semeraro, J. Montomoli, V. Bellini, O. Piazza, and E. Bignami, "The breakthrough of large language models release for medical applications: 1-year timeline and perspectives," *Journal of Medical Systems*, 2024, doi: 10.1007/s10916-024-02045-3.

[18] O. K. Gargari and G. Habibi, "Enhancing medical AI with retrieval-augmented generation: A mini narrative review," *Digital Health*, 2025, doi: 10.1177/20552076251337177.

[19] S. Kresevic, M. Giuffré, M. Ajčević, A. Accardo, L. Crocè, and D. Shung, "Optimization of hepatological clinical guidelines interpretation by large language models: A retrieval augmented generation-based framework," *NPJ Digital Medicine*, 2024, doi: 10.1038/s41746-024-01091-y.

[20] R. Li *et al.*, "Enhancing pulmonary disease prediction using large language models with feature summarization and hybrid retrieval-augmented generation: Multicenter methodological study based on radiology report," *Journal of Medical Internet Research*, 2025, doi: 10.2196/72638.

[21] D. Baur, J. Ansorg, C.-E. Heyde, and A. Voelker, "Development and evaluation of a retrieval-augmented generation chatbot for orthopedic and trauma surgery patient education: Mixed-methods study," *JMIR AI*, 2025, doi: 10.2196/75262.

[22] X. Ge, S. Murtaza, A. Cortez, and H. Alemzadeh, "Expert-guided prompting and retrieval-augmented generation for emergency medical service question answering," in *Proc. AAAI Conf. Artificial Intelligence*, 2025, doi: 10.48550/arXiv.2511.10900.

[23] Z. Wang, D. Wu, Q. Nguyen, and K. Wang, "Integrating chain-of-thought and retrieval augmented generation enhances rare disease diagnosis from clinical notes," *Medicine Bulletin*, 2026, doi: 10.1002/mdb2.70033.

[24] Y. Gao *et al.*, "Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study," *JMIR AI*, 2024, doi: 10.2196/58670.

[25] C. Park, H. Lee, S. Lee, and O. Jeong, "Synergistic joint model of knowledge graph and LLM for enhancing XAI-based clinical decision support systems," *Mathematics*, 2025, doi: 10.3390/math13060949.

[26] J. Bélisle-Pipon, "Why we need to be careful with LLMs in medicine," *Frontiers in Medicine*, 2024, doi: 10.3389/fmed.2024.1495582.

[27] Q. Jia *et al*., "Implicit bias in safety-aligned large language models: A multi-faceted evaluation of clinical decision-making and health equity," *PLoS ONE*, 2026, doi: 10.1371/journal.pone.0348819.

[28] I. Papagiannopoulos, H. Koutalidis, P. Rempi, C. Ntanos, and D. Askounis, "Comparison of explainability methods for hallucination analysis in LLMs," *Open Research Europe*, 2025, doi: 10.12688/openreseurope.20839.1.

[29] M. Griot, C. Hemptinne, J. Vanderdonckt, and D. Yuksel, "Large language models lack essential metacognition for reliable medical reasoning," *Nature Communications*, 2025, doi: 10.1038/s41467-024-55628-6.

[30] A. Bhatia, R. Kumar, and G. Manteghi, "Formulating reliable deep acquisition hypotheses for medical diagnosis utilizing explainable AI," *International Journal of Advanced Medical Sciences and Technology*, 2025, doi: 10.54105/ijamst.f3050.05061025.

[31] S. Y. Kim, D. H. Kim, M. J. Kim, H. Ko, and O. Jeong, "XAI-based clinical decision support systems: A systematic review," *Applied Sciences*, 2024, doi: 10.3390/app14156638.

[32] E. Grzegorzak *et al*., "Explainable artificial intelligence in healthcare: From algorithmic transparency to trust and social acceptance in clinical practice," *International Journal of Innovative Technologies in Social Science*, 2026, doi: 10.31435/ijitss.1(49).2026.4676.

[33] O. Ozdemir, "Explainable AI (XAI) in healthcare: Bridging the gap between accuracy and interpretability," *Journal of Science, Technology and Engineering Research*, 2024, doi: 10.64206/0z78ev10.

[34] M. Ahmed, O. Okesanya, M. Oweidat, Z. Othman, S. Musa, and D. E. Lucero-Prisno III, "The ethics of data mining in healthcare: Challenges, frameworks, and future directions," *BioData Mining*, 2025, doi: 10.1186/s13040-025-00461-w.

[35] J. Zhang and Z.-M. Zhang, "Ethics and governance of trustworthy medical artificial intelligence," *BMC Medical Informatics and Decision Making*, 2023, doi: 10.1186/s12911-023-02103-9.

[36] H. Raposo and J. Umer, "An uncertainty-aware multimodal deep learning framework for robust cancer diagnosis under dataset shift," *Social Science Research Network*, 2026, doi: 10.2139/ssrn.6512918.

[37] H. Koo, "Towards robust and fair next visit diagnosis prediction under noisy clinical notes with large language models," *arXiv*, 2025, doi: 10.48550/arXiv.2511.18393.

[38] S. F. Begum and F. Sultana, "Explainable and trustworthy generative AI: A framework for interpretable large language models in high-stakes decision systems," *International Journal of Advanced Research in Science, Communication and Technology*, 2026, doi: 10.48175/ijarsct-31102.

[39] P. Goktas and A. E. Grzybowski, "Shaping the future of healthcare: Ethical clinical challenges and pathways to trustworthy AI," *Journal of Clinical Medicine*, 2025, doi: 10.3390/jcm14051605.

[40] S. Tsioutra and A. Davalas, “Preparing healthcare organizations for compliance with the European AI Act,” *International Journal of Social Science and Economic Research*, 2026, doi: 10.46609/ijsser.2026.v11i02.035.

[41] N. Rajashekar *et al*., “Human-algorithmic interaction using a large language model-augmented artificial intelligence clinical decision support system,” in *Proc. Int. Conf. Human Factors in Computing Systems*, 2024, doi: 10.1145/3613904.3642024.

[42] C. U. Ogdu, S. Gurbuz, M. Karakose, and E. Hanoğlu, “Medical implications of LLM based clinical decision support systems in healthcare,” in *Int. Conf. Information Technology*, 2025, doi: 10.1109/IT64745.2025.10930252.

[43] X. Chen *et al*., “EyeGPT for patient inquiries and medical education: Development and validation of an ophthalmology large language model,” *Journal of Medical Internet Research*, 2024, doi: 10.2196/60063.

[44] J. Ge *et al*., “Development of a liver disease–specific large language model chat interface using retrieval-augmented generation,” *Hepatology*, 2024, doi: 10.1097/hep.0000000000000834.

[45] M. G. G. Pérez *et al*., “Hematobot: Retrieval-augmented large language model-based assistants for therapy, toxicity and frailty management in multiple myeloma by the Spanish Society of Hematology and Hemotherapy (SEHH),” *Blood*, 2025, doi: 10.1182/blood-2025-6115.

[46] S. Jia, S. Bit, V. H. Jasodanand, Y. Liu, and V. Kolachalama, “Agentic memory-augmented retrieval and evidence grounding for medical question-answering tasks,” *medRxiv*, 2025, doi: 10.1101/2025.08.06.25333160.

[47] A. P. D. Ishadya, A. Withanaarachchi, and S. Jayalal, “Mitigating bias and hallucinations in LLMs through prompt engineering and knowledge-grounded approaches in healthcare domain: A systematic literature review,” in *2025 Int. Conf. Advances in Technology and Computing (ICATC)*, 2025, doi: 10.1109/ICATC68823.2025.11407841.

[48] Kumar, Anuj. "Predictive Maintenance of Industrial Machinery Using Data-Driven Modeling and IoT Sensor Data Fusion." International Journal of Engineering Research and Application, Vol. 2, Issue 6, 2012, pp. 1712–1719